

# The cavity method for analysis of large-scale penalized regression

Mohammad Ramezanali,<sup>1,\*</sup> Partha P. Mitra,<sup>2,†</sup> and Anirvan M. Sengupta<sup>1,3,‡</sup>

<sup>1</sup>*Department of Physics and Astronomy, Rutgers University,  
136 Frelinghuysen Rd, Piscataway, NJ 08854 USA*

<sup>2</sup>*Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11734 USA*

<sup>3</sup>*Center for Quantitative Biology, Rutgers University,  
110 Frelinghuysen Rd, Piscataway, NJ 08854 USA*

Penalized regression methods aim to retrieve reliable predictors among a large set of putative ones from a limited amount of measurements. In particular, penalized regression with singular penalty functions is important for sparse reconstruction algorithms. For large-scale problems, these algorithms exhibit sharp phase transition boundaries where sparse retrieval breaks down. Large optimization problems associated with sparse reconstruction have been analyzed in the literature by setting up corresponding statistical mechanical models at a finite temperature. Using replica method for mean field approximation, and subsequently taking a zero temperature limit, this approach reproduces the algorithmic phase transition boundaries. Unfortunately, the replica trick and the non-trivial zero temperature limit obscure the underlying reasons for the failure of a sparse reconstruction algorithm, and of penalized regression methods, in general. In this paper, we employ the “cavity method” to give an alternative derivation of the mean field equations, working directly in the zero-temperature limit. This derivation provides insight into the origin of the different terms in the self-consistency conditions. The cavity method naturally involves a quantity, the average local susceptibility, whose behavior distinguishes different phases in this system. This susceptibility can be generalized for analysis of a broader class of sparse reconstruction algorithms.

PACS numbers: 61.50.Ah

Keywords: cavity method; penalized regression; sparsity; compressed sensing

## I. INTRODUCTION

In traditional statistics, we are given a sample of random observations much larger than the number of unknown parameters being estimated. However, during the last couple of decades, data collection has become much more automatic and much more extensive. As a result, the limit of large number of unknown parameters appears in a variety of fields, ranging from communication technology to business informatics to systems biology, posing challenges to the classical statistical paradigm. Many methods for the selection of important variables, proposed within the statistics literature, are combinatorial in nature [1]. The explosion of number of possibilities, when the number of unknown variables are comparable to the sample size, places a huge computational burden on the more principled methods, severely limiting their applicability to large data sets. Practical variable selection procedures drastically limit the search space, but they are often greedy in nature and potentially unreliable.

Variable selection via penalized regression [2, 3] has gained wide appeal, owing to these concerns regarding combinatorial methods. One common setup is a linear model,  $\mathbf{y} = \mathbf{H}\mathbf{x}_0 + \boldsymbol{\zeta}$ , where it is assumed that  $\mathbf{H}$  is a known  $M \times N$  measurement matrix,  $\mathbf{x}_0$  is an unknown  $N$ -dimensional signal vector, possibly sparse, and the

vector  $\boldsymbol{\zeta}$  models the measurement noise. The task is to reconstruct  $\mathbf{x}_0$  from  $\mathbf{y}$  and  $\mathbf{H}$ , utilizing, in principle, the knowledge about the statistics of the signal and of noise.

The general form of penalized methods for standard linear model, in practice, does not require detailed information about statistics of  $\mathbf{x}_0$  and  $\boldsymbol{\zeta}$ . The task reduces to minimizing the following objective function

$$\hat{\mathbf{x}}(\lambda\sigma^2) = \arg \min_{\mathbf{x}} \left\{ \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{H}\mathbf{x})^2 + \lambda V(\mathbf{x}) \right\}. \quad (1)$$

where  $\lambda\sigma^2$  is a non-negative parameter deciding relative weight between quadratic loss function and the penalty function  $V$  in Eq. (1). The penalty function  $V$  is chosen to suppress effects of noise on estimation as well as to regularize potentially ‘soft’ directions in the loss function. A popular class of penalties are of the form  $\sum_a |x_a|^m$ . The choice  $m = 2$  or  $\ell_2$  penalty is known as Ridge regression [4], a convex penalty function which has been shown to give more accurate predications, when sparsity of the solution is not important. Choosing  $m \rightarrow 0$  yields the so-called  $\ell_0$  penalty which controls sparsity directly but leads to a non-convex optimization problem. Finally,  $m = 1$  or  $\ell_1$  penalty, has the benefits of being a convex function as well as of promoting sparsity [2, 5].

When the vector  $\mathbf{x}$  has  $K$  non-zero components, with  $K \leq M$  and  $M \leq N$ , there are many rigorous results on the performance of sparse reconstruction based on  $\ell_1$ -penalization [6, 7]. As a result of the computational attractiveness of convex optimization and of the performance guarantees which reach optimality under certain conditions,  $\ell_1$ -penalization has become a popular

\* mrr@physics.rutgers.edu

† mitra@cshl.edu

‡ anirvans@physics.rutgers.edu

method, particularly in the context of so-called Compressed Sensing [7]. In the asymptotic limit  $M, N, K \rightarrow \infty$ , there is a phase transition in noiseless sparse recovery [8, 9], with constrained optimization of the  $\ell_1$  penalty leading to perfect reconstruction in the ‘good’ phase.

As one can imagine, the statistical physics of disordered systems offers powerful tools to understand this phase transition in the asymptotic limit. In a series of works using the *replica method*, several investigators have studied the performance of  $\ell_1$ -penalized regression and the algorithmic phase transition [10–14]. In comparison to rigorous results providing guarantees [6], the statistical mechanics approach enables a detailed analysis of the behavior of the optimization algorithms near the region of failure. On the other hand, the replica trick is applied to a finite temperature system. The non-trivial zero temperature limit hides the essential role of local susceptibility in deciding the robustness of performance of the algorithms. Also, the derivation of the self-consistency conditions obscures the subtlety of certain aspects of this mean field theory.

An alternative to the replica trick, with the number of replicas ‘mysteriously’ tending to zero, is to use the *cavity method* [15, 16]. It is a direct approach to mean field theory, originally designed for understanding the nature of ground states of certain spin glass models. The method has since been applied to a wider class of problems including algorithmic phase transitions, some examples being the satisfiability problem [17, 18] and Hopfield neural networks [19]. The cavity method leads to the same results as obtained by replica trick [16] for spin glass mean field theory and is closely related to the message-passing algorithm in graphical models [20]. We find that for the problem at hand, the cavity method provides better insight in comparison to the replica formalism and has the potential to lead to substantially better sparse reconstruction algorithms.

The layout of this paper is as follows. In the next two sections, we briefly review a finite noise/finite temperature formulation of the problem and the replica approach to the mean field theory, facilitating comparison with our analysis via the cavity method. Readers familiar with these results, or those solely interested in the cavity method, could skim through these sections just familiarizing themselves with the notation. In Sec. IV, we introduce a susceptibility matrix associated with this problem and then provide an outline of the derivation of the self-consistent mean field equations via a two-step cavity method working directly at zero temperature. This approach is not only different from the replica method but also from the analysis based on iterations in a message-passing algorithm [21, 22]. We end by illustrating how our cavity mean field picture relate to success and to failure of sparse reconstruction in medium sized  $\ell_1$ -penalized problems in Sec. V. The appendix has additional technical details of the derivation, along with the cavity approach worked out for finite temperature.

## II. STATISTICAL MECHANICS FORMULATION

Here, we set up the general framework for investigating the regularized least-squares reconstruction algorithms. We assume that the data  $\mathbf{y} = \mathbf{H}\mathbf{x}_0 + \boldsymbol{\zeta}$  are generated by a probability distribution  $p(\mathbf{y}|\mathbf{x}_0, H)$ , given an (unknown) sparse signal  $\mathbf{x}_0$  and a (known) matrix  $\mathbf{H}$ , and an (unknown) Gaussian noise vector  $\boldsymbol{\zeta}$  whose components are i.i.d. samples from  $\mathcal{N}(0, \sigma_\zeta^2)$ . The vector  $\mathbf{x}_0$  is considered to be a random sample from a distribution  $P_0(\mathbf{x}_0) = \prod_a p_0(x_{a0})$ .

Although, in general, the probability distribution of  $\mathbf{H}$ ,  $\mathcal{P}(\mathbf{H})$ , could be a non-Gaussian distribution, at this point we consider it to be a multivariate Gaussian distribution. The element-wise mean and the covariance matrix entries are given by

$$[H_{ia}]^{\text{av}} = 0 \quad (2)$$

$$[H_{ia}H_{jb}]^{\text{av}} = \frac{1}{M}\delta_{ij}\delta_{ab}. \quad (3)$$

We study the performance of an estimator of  $\mathbf{x}_0$ , namely the location  $\hat{\mathbf{x}}$  of the minimum of a cost function

$$\mathcal{E}_0(\mathbf{x}) = \frac{(\mathbf{y} - \mathbf{H}\mathbf{x})^2}{2\sigma^2} + V(\mathbf{x}). \quad (4)$$

We can reformulate the cost minimization as a statistical mechanics problem where the cost function will play the role of energy. We assume the penalty/potential term  $V(\mathbf{x})$  is such that there is a unique minimum of  $\mathcal{E}_0$ . Note that  $\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \mathcal{E}_0(\mathbf{x})$  depends on  $\mathbf{y}, \mathbf{H}$ , meaning that it can be written as a function  $\hat{\mathbf{x}} = \mathbf{g}(\mathbf{x}_0, \mathbf{H}, \boldsymbol{\zeta})$ , using the fact that  $\mathbf{y} = \mathbf{H}\mathbf{x}_0 + \boldsymbol{\zeta}$ . We have set up an ensemble of problem instances by specifying the probability distribution of the variables  $\mathbf{x}_0, \mathbf{H}, \boldsymbol{\zeta}$ , so that we could study the performance of the estimator over this ensemble. In particular, it will be useful to extract moments of the distribution of the parameter estimation error  $\hat{\mathbf{x}} - \mathbf{x}_0$ .

In order to make a connection between the optimizations problem and statistical mechanics, one could choose a probability distribution of  $\mathbf{x}$  parametrized by  $\beta$ , playing the role of inverse temperature,

$$\begin{aligned} p_\beta(\mathbf{x}|\mathbf{y}, \mathbf{H}) &= \frac{1}{Z} \exp(-\beta \mathcal{E}_0(\mathbf{x})) \\ &= \frac{1}{Z(\beta, \mathbf{y}, \mathbf{H})} \exp\left\{-\beta \left(\frac{(\mathbf{y} - \mathbf{H}\mathbf{x})^2}{2\sigma^2} + V(\mathbf{x})\right)\right\} \end{aligned} \quad (5)$$

with the normalization factor  $Z = Z(\beta, \mathbf{y}, \mathbf{H})$ , known as

the partition function, given by

$$\begin{aligned} Z(\beta, \mathbf{y}, \mathbf{H}) &= \int d^N \mathbf{x} \frac{1}{Z} \exp(-\beta \mathcal{E}_0(\mathbf{x})) \\ &= \int d^N \mathbf{x} \exp \left\{ -\beta \left( \frac{(\mathbf{y} - \mathbf{H}\mathbf{x})^2}{2\sigma^2} + V(\mathbf{x}) \right) \right\}. \end{aligned} \quad (6)$$

If we send  $\beta$  to  $\infty$ , equivalent to sending the temperature to zero, the probability gets concentrated at the minimum of the cost/energy function. Keep in mind that we define  $\beta$  to be dimensionless.

We will consider averages of functions of the form  $\mathcal{O}(\mathbf{x}, \mathbf{x}_0)$  containing both the original sparse signal and the variable related to the estimate. The average of the function  $\mathcal{O}(\mathbf{x}, \mathbf{x}_0)$  over the distribution  $p_\beta(\mathbf{x}|\mathbf{y}, \mathbf{H})$  is given by

$$\langle \mathcal{O}(\mathbf{x}, \mathbf{x}_0) \rangle = \frac{\int d^N \mathbf{x} \mathcal{O}(\mathbf{x}, \mathbf{x}_0) \exp \left\{ -\beta \frac{(\mathbf{y} - \mathbf{H}\mathbf{x})^2}{2\sigma^2} - \beta V(\mathbf{x}) \right\}}{\int d^N \mathbf{x} \exp \left\{ -\beta \frac{(\mathbf{y} - \mathbf{H}\mathbf{x})^2}{2\sigma^2} - \beta V(\mathbf{x}) \right\}}. \quad (7)$$

This ‘thermal’ average, represented by  $\langle \dots \rangle$ , depends on the random variables  $\mathbf{x}_0$ ,  $\mathbf{H}$  and  $\zeta$ . Note that in the limit  $\beta \rightarrow \infty$ , this average should become  $f(\hat{\mathbf{x}}, \mathbf{x}_0)$ , for continuous  $f$ . Averaging the result of this calculation over the random instances of  $\mathbf{x}_0$ ,  $\mathbf{H}$  and  $\zeta$  is a technical challenge related to quenched averages in disordered systems [16].

The function  $\mathcal{O}(\mathbf{x}, \mathbf{x}_0) = \frac{1}{N}(\mathbf{x} - \mathbf{x}_0)^2$  plays an important role in our analysis. Its average corresponds to the mean squared estimation error:

$$\text{MSE} \equiv \frac{1}{N} \sum_{a=1}^N [\langle x_a - x_{a0} \rangle^2]_{\mathbf{x}_0, \mathbf{H}, \zeta}^{\text{av}} = \frac{1}{N} [\langle (\mathbf{x} - \mathbf{x}_0)^2 \rangle]_{\mathbf{x}_0, \mathbf{H}, \zeta}^{\text{av}}. \quad (8)$$

We will use  $[\dots]_{\text{vars}}^{\text{av}}$  to denote quenched averages, with the relevant quenched variables indicated in the subscript, when necessary. We use the notation  $\mathbf{u} = \mathbf{x} - \mathbf{x}_0$  to indicate the estimation error vector. The size of the vector  $\mathbf{u}$  provides a measure of the inaccuracy of the reconstruction.

In the context of penalized regression, the penalty function is often chosen to be a sum of potentials involving single variables, namely,  $V(\mathbf{x}) = \sum_a U(x_a)$ . We will focus on  $V(\mathbf{x})$  of this nature. An important special case for example is in compressed sensing with sparsity promoting regularizing potential  $U(x) = \lambda|x|$ . In this paper, we would mostly restrict ourselves to the noiseless case,  $\zeta = 0$ , although the same methods could be used to analyze the noisy case as well.

For  $\zeta = 0$ , we will be interested in the result of the constrained optimization problem of minimizing  $V(\mathbf{x})$  subject to the constraint  $\mathbf{y} = \mathbf{H}\mathbf{x}$ . In the  $M, N \rightarrow \infty$  limit, this problem may exhibit a phase transition from a perfect reconstruction phase to an error-prone phase, with the MSE, mentioned above, as the order parameter. This constrained optimization could be studied in more than one equivalent ways. After taking  $\beta \rightarrow \infty$  limit, we will take the route  $\sigma \rightarrow 0$  to enforce the equality  $\mathbf{y} = \mathbf{H}\mathbf{x}$ .

### III. REPLICA APPROACH

In this section, we review the replica approach to the problem [13, 14], presenting the mean field equations in terms of a distribution of asymptotically independent single-variable problems with a set of self-consistency conditions. In order to calculate quantities like the MSE, we need to compute quenched averages of the form  $[\langle \mathcal{O}(\mathbf{x}, \mathbf{x}_0) \rangle]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}}$ , which is complicated by the presence of the denominator in Eq. (7). Formally, the denominator is handled by introducing  $n$  non-interacting replicas of the system and taking  $n \rightarrow 0$ , as shown below. In the noiseless case,  $\mathcal{E}_0(\mathbf{x})$  depends on  $\mathbf{x}$  as well as on  $\mathbf{x}_0, H$ . To emphasize those additional dependences, we write  $\mathcal{E}_0(\mathbf{x})$  as  $\mathcal{E}_0(\mathbf{x}_\mu, \mathbf{x}_0, \mathbf{H})$  in the next few equations.

$$\begin{aligned} \langle \mathcal{O}(\mathbf{x}, \mathbf{x}_0) \rangle_{\mathbf{x}} &= \frac{\int d^N \mathbf{x} \mathcal{O}(\mathbf{x}, \mathbf{x}_0) \exp(-\beta \mathcal{E}_0(\mathbf{x}, \mathbf{x}_0, \mathbf{H}))}{\int d^N \mathbf{x} \exp(-\beta \mathcal{E}_0(\mathbf{x}, \mathbf{x}_0, \mathbf{H}))} \\ &= \lim_{n \rightarrow 0} \left( \int d^N \mathbf{x} \exp(-\beta \mathcal{E}_0(\mathbf{x}, \mathbf{x}_0, \mathbf{H})) \right)^{n-1} \\ &\quad \int d^N \mathbf{x} \mathcal{O}(\mathbf{x}, \mathbf{x}_0) \exp(-\beta \mathcal{E}_0(\mathbf{x}, \mathbf{x}_0, \mathbf{H})) \\ &= \lim_{n \rightarrow 0} \int \mathcal{O}(\mathbf{x}_1, \mathbf{x}_0) \prod_{\mu=1}^n \{ d^N \mathbf{x}_\mu \exp(-\beta \mathcal{E}_0(\mathbf{x}_\mu, \mathbf{x}_0, \mathbf{H})) \}. \end{aligned} \quad (9)$$

Averaging over the quenched variables  $\mathbf{x}_0$  and  $H$ , we get

$$\begin{aligned} [\langle \mathcal{O}(\mathbf{x}, \mathbf{x}_0) \rangle_{\mathbf{x}}]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} &= \lim_{n \rightarrow 0} \left[ \int \prod_{\mu=1}^n \{ d^N \mathbf{x}_\mu \} \mathcal{O}(\mathbf{x}_1, \mathbf{x}_0) \exp(-\beta \sum_{\mu} \mathcal{E}_0(\mathbf{x}_\mu, \mathbf{x}_0, \mathbf{H})) \right]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}}. \end{aligned} \quad (10)$$

Using  $\mathbf{y} = \mathbf{H}\mathbf{x}_0$  in the noiseless case, the energy function for the  $n$ -th replica would be

$$\begin{aligned} \mathcal{E}_0(\mathbf{x}_\mu, \mathbf{x}_0, \mathbf{H}) &= \frac{(\mathbf{y} - \mathbf{H}\mathbf{x}_\mu)^2}{2\sigma^2} + V(\mathbf{x}_\mu) \\ &= \frac{(\mathbf{H}\mathbf{x}_0 - \mathbf{H}\mathbf{x}_\mu)^2}{2\sigma^2} + V(\mathbf{x}_\mu) = \frac{\mathbf{H}\mathbf{u}_\mu^2}{2\sigma^2} + V(\mathbf{u}_\mu + \mathbf{x}_0), \end{aligned} \quad (11)$$

rewritten in terms of the error variables  $\mathbf{u}_\mu = \mathbf{x}_\mu - \mathbf{x}_0$ ,  $\mu = 1, \dots, n$ . Thus, we are interested in average quantities in the replicated ensemble whose partition functions is given by

$$\begin{aligned} [Z^n]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} &= \left[ \int \prod_{\mu=1}^n d\mathbf{u}_\mu \exp \left[ -\beta \left\{ \sum_{\mu=1}^n \frac{(\mathbf{H}\mathbf{u}_\mu)^2}{2\sigma^2} + V(\mathbf{u}_\mu + \mathbf{x}_0) \right\} \right] \right]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}}. \end{aligned} \quad (12)$$

In order to average over  $\mathcal{P}(\mathbf{H})$ , the only quantity that needs to be computed is

$$\begin{aligned} & \left[ \exp \left( - \sum_{\mu=1}^n \frac{\beta}{2\sigma^2} (\mathbf{H}\mathbf{u}_\mu)^2 \right) \right]_{\mathbf{H}}^{\text{av}} \\ &= \frac{1}{Z_0} \int d\mathbf{H} \exp \left[ - \frac{M}{2} \text{Tr}(\mathbf{H}^\top \mathbf{H}) - \frac{\beta}{2\sigma^2} \sum_{\mu=1}^n \mathbf{u}_\mu^\top \mathbf{H}^\top \mathbf{H} \mathbf{u}_\mu \right] \\ &= \left[ \frac{1}{\det \left( \mathbf{I}_N + \frac{\beta}{\alpha\sigma^2} \sum_{\mu=1}^n \frac{1}{N} \mathbf{u}_\mu \mathbf{u}_\mu^\top \right)} \right]^{-M/2} \end{aligned} \quad (13)$$

where  $Z_0$  is the normalization term for the Gaussian distribution of  $\mathbf{H}$ , and  $\alpha = M/N$  is the sampling ratio. We can look at the  $\phi = \mathbf{u}\mathbf{u}^\top$  where  $\mathbf{u}$  is an  $N$  by  $n$  matrix,  $[\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n]$ . We write the singular value decomposition (SVD) for  $\mathbf{u}$  as

$$\mathbf{u} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top \quad (14)$$

where  $\mathbf{\Lambda}$  denotes the  $n$  by  $n$  matrix with  $n$  non-zero singular values equals  $\lambda_\mu$ . Therefore, the eigenvalues of  $\phi$  are simply the square of these singular values. Changing the variable from  $\mathbf{u}$  in (13) to  $\mathbf{U}$ ,  $\mathbf{V}$ , and  $\mathbf{\Lambda}$ :

$$\det \left( \mathbf{I}_N + \frac{\beta}{\alpha\sigma^2} \frac{1}{N} \mathbf{u}\mathbf{u}^\top \right) = \prod_{\mu=1}^n \left( 1 + \frac{\beta}{\alpha\sigma^2} \frac{1}{N} \lambda_\mu^2 \right). \quad (15)$$

The right hand side of Eq. (15) can be written as  $\det \left( \mathbf{I}_n + \frac{\beta}{\alpha\sigma^2} \mathbf{Q} \right)$  with the elements of the  $n \times n$  matrix  $\mathbf{Q}$  are defined by  $\mathbf{Q} = \frac{1}{N} \mathbf{u}^\top \mathbf{u}$  which have the same  $\lambda_\mu^2$  eigenvalues. Therefore, we can rewrite Eq. (12) as

$$\begin{aligned} & [Z^n]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} \\ &= \left[ \int \prod_{\mu=1}^n d\mathbf{u}_\mu \exp \left[ - \frac{M}{2} \text{Tr} \log \left( \mathbf{I}_n + \frac{\beta}{\alpha\sigma^2} \mathbf{Q} \right) + \sum_{\mu=1}^n V(\mathbf{u}_\mu + \mathbf{x}_0) \right] \right]_{\mathbf{x}_0} \end{aligned} \quad (16)$$

Using the Fourier representation of the  $\delta$  function

$$\delta(\mathbf{u}_\mu^\top \mathbf{u}_\nu - N Q_{\mu\nu}) = \frac{1}{2\pi} \int dR_{\mu\nu} \exp(-iR_{\mu\nu}(\mathbf{u}_\mu^\top \mathbf{u}_\nu - N Q_{\mu\nu})) \quad (17)$$

and inserting this delta function with an integral over  $\mathbf{Q}_{\mu\nu}$  in Eq. (12), we get

$$[Z^n]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} = \int \prod_{\mu \leq \nu} dQ_{\mu\nu} dR_{\mu\nu} \exp[-S(\mathbf{Q}, \mathbf{R})] \quad (18)$$

$$\begin{aligned} S[\mathbf{Q}, \mathbf{R}] &= \frac{M}{2} \text{Tr} \log \left( \mathbf{I}_n + \frac{\beta}{\alpha\sigma^2} \mathbf{Q} \right) - iN \text{Tr}(\mathbf{R}\mathbf{Q}) \\ &- \log \left[ \int \prod_{\mu=1}^n d\mathbf{u}_\mu \exp \left[ -i \sum_{\mu, \nu} R_{\mu\nu} \mathbf{u}_\mu^\top \mathbf{u}_\nu + \sum_{\mu} V(\mathbf{u}_\mu + \mathbf{x}_0) \right] \right]_{\mathbf{x}_0}^{\text{av}} \end{aligned} \quad (19)$$

This integral over  $\mathbf{Q}, \mathbf{R}$  can be evaluated using the saddle point method [13, 14] when  $M, N \rightarrow \infty$ , holding  $\alpha = \frac{M}{N}$

fixed. The saddle point  $\mathbf{Q} = \bar{\mathbf{Q}}, \mathbf{R} = -i\bar{\mathbf{R}}$  satisfies the conditions:

$$\bar{Q}_{\mu\nu} = \frac{1}{N} \langle \mathbf{u}_\mu^\top \mathbf{u}_\nu \rangle \quad (20)$$

$$\bar{\mathbf{R}} = \frac{\beta}{2\sigma^2} \left[ \mathbf{I}_n + \frac{\beta}{\alpha\sigma^2} \mathbf{Q} \right]^{-1} \quad (21)$$

obtained by differentiating  $S(\mathbf{Q}, \mathbf{R})$  with respect to the elements of  $\mathbf{Q}, \mathbf{R}$ . The expectation  $\langle \mathbf{u}_\mu^\top \mathbf{u}_\nu \rangle$  depends on  $\bar{\mathbf{R}}$  via

$$\langle \mathbf{u}_\mu^\top \mathbf{u}_\nu \rangle = \beta \frac{\partial F(\bar{\mathbf{R}})}{\partial \bar{R}_{\mu\nu}} \quad (22)$$

with  $\exp\{-\beta F(\bar{\mathbf{R}})\}$

$$= \left[ \int \prod_{\mu=1}^n \{d^N \mathbf{u}_\mu\} \exp \left[ - \sum_{\mu, \nu} \bar{R}_{\mu\nu} \mathbf{u}_\mu^\top \mathbf{u}_\nu - \beta \sum_{\mu} V(\mathbf{u}_\mu + \mathbf{x}_0) \right] \right]_{\mathbf{x}_0}^{\text{av}}. \quad (23)$$

If  $U(x)$  is a convex function, we expect a unique state and a replica symmetric solution for  $\mathbf{Q}, \mathbf{R}$ . This implies  $\bar{Q}_{\mu\nu} = (Q - q)\delta_{\mu\nu} + q$  and  $\bar{R}_{\mu\nu} = (R - r)\delta_{\mu\nu} + r$ . With that ansatz,

$$\begin{aligned} & \int \prod_{\mu=1}^n \{d^N \mathbf{u}_\mu\} \exp \left[ - \sum_{\mu, \nu} \bar{R}_{\mu\nu} \mathbf{u}_\mu^\top \mathbf{u}_\nu - \beta \sum_{\mu} V(\mathbf{u}_\mu + \mathbf{x}_0) \right] \\ &= \int \prod_{\mu=1}^n \{d^N \mathbf{u}_\mu\} \exp \left[ - (R - r) \sum_{\mu} \mathbf{u}_\mu^2 \right. \\ &\quad \left. - r \left( \sum_{\mu} \mathbf{u}_\mu \right)^2 - \beta \sum_{\mu} V(\mathbf{u}_\mu + \mathbf{x}_0) \right] \\ &= \int \frac{d^N \boldsymbol{\xi}}{(2\pi\sigma_\xi^2)^{N/2}} \exp\left(-\frac{\boldsymbol{\xi}^2}{2\sigma_\xi^2}\right) \int \prod_{\mu=1}^n \{d^N \mathbf{u}_\mu\} \\ &\quad \exp \left[ - \frac{\beta}{2\sigma_{\text{eff}}^2} \sum_{\mu} \mathbf{u}_\mu^2 + \frac{\beta}{\sigma_{\text{eff}}^2} \boldsymbol{\xi}^\top \left( \sum_{\mu} \mathbf{u}_\mu \right) - \beta \sum_{\mu} V(\mathbf{u}_\mu + \mathbf{x}_0) \right] \end{aligned} \quad (24)$$

identifying  $R - r \equiv \frac{\beta}{2\sigma_{\text{eff}}^2}$  and  $r \equiv -\frac{\beta^2 \sigma_\xi^2}{2\sigma_{\text{eff}}^4}$ . We have used

$$\begin{aligned} & \int \frac{d^N \boldsymbol{\xi}}{(2\pi\sigma_\xi^2)^{N/2}} \exp\left(-\frac{\boldsymbol{\xi}^2}{2\sigma_\xi^2}\right) \exp\left[\frac{\beta}{\sigma_{\text{eff}}^2} \boldsymbol{\xi}^\top \left( \sum_{\mu} \mathbf{u}_\mu \right)\right] \\ &= \exp\left[\frac{\beta^2 \sigma_\xi^2}{2\sigma_{\text{eff}}^4} \left( \sum_{\mu} \mathbf{u}_\mu \right)^2\right] \end{aligned} \quad (25)$$

to decouple the item replica coupling in the  $(\sum_{\mu} \mathbf{u}_\mu)^2$  term, at the cost of introducing another quenched variable  $\boldsymbol{\xi}$ . Note that we require  $R - r > 0$  and  $r < 0$  for this approach to work. These inequalities follow from (21) and from  $Q - q > 0$  and  $q > 0$ . The conditions on  $Q$  and  $q$  would be obvious once we look at interpretation of these quantities described below.

For  $V(\mathbf{x}) = \sum_a U(x_a)$  we can simplify further. Remembering that we also need to do the quenched average over

$\mathbf{x}_0$ ,

$$\begin{aligned}
& \left[ \int \prod_{\mu=1}^n \{d^N \mathbf{u}_\mu\} \exp \left[ - \sum_{\mu,\nu} \bar{R}_{\mu\nu} \mathbf{u}_\mu^\top \mathbf{u}_\nu - \beta \sum_{\mu} V(\mathbf{u}_\mu + \mathbf{x}_0) \right] \right]_{\mathbf{x}_0}^{\text{av}} \\
&= \left[ \int \prod_{\mu=1}^n \{d^N \mathbf{u}_\mu\} \exp \left[ - \beta \left\{ \frac{1}{2\sigma_{\text{eff}}^2} \sum_{\mu} (\mathbf{u}_\mu^2 - \xi^\top \mathbf{u}_\mu) \right. \right. \right. \\
&\quad \left. \left. \left. + \sum_{\mu} V(\mathbf{u}_\mu + \mathbf{x}_0) \right\} \right] \right]_{\xi, \mathbf{x}_0}^{\text{av}} \\
&= \prod_a \left[ \int \prod_{\mu=1}^n \{du_{\mu a}\} \exp \left[ - \beta \left\{ \frac{1}{2\sigma_{\text{eff}}^2} \sum_{\mu} (u_{\mu a}^2 - \xi_a u_{\mu a}) \right. \right. \right. \\
&\quad \left. \left. \left. + \sum_{\mu} U(u_{\mu a} + x_{0a}) \right\} \right] \right]_{\xi_a, x_{0a}}^{\text{av}} \quad (26)
\end{aligned}$$

Thus, in the saddle point approximation, each of the  $N$  components of  $\mathbf{u}$  become effectively independent and the saddle point conditions reduce to a self-consistent problem for each component  $a = 1, \dots, N$ . Since this self-consistent problem is similar for each index, we suppress the subscript  $a$  in  $u_{\mu a}$  and in  $x_{0a}$ . For each  $a$ , we have the integral of the form

$$\left[ \int \prod_{\mu=1}^n du_{\mu} \exp \left[ - \beta \left\{ \frac{1}{2\sigma_{\text{eff}}^2} \sum_{\mu} (u_{\mu}^2 - \xi u_{\mu}) + \sum_{\mu} U(u_{\mu} + x_0) \right\} \right] \right]_{\xi, x_0}^{\text{av}}$$

The replica problem corresponds to a single variable  $u$  follows the effective distribution

$$P_{\text{eff}}(u | x_0, \xi) = \frac{1}{Z(x_0, \xi)} e^{-\beta \mathcal{E}_{\text{eff}}(u; x_0, \xi)}, \quad (27)$$

with an effective mean-field Hamiltonian

$$\mathcal{E}_{\text{eff}}(u; x_0, \xi) = \frac{1}{2\sigma_{\text{eff}}^2} (u^2 - 2\xi u) + U(u + x_0) \quad (28)$$

which depends on two quenched variables  $x_0$  and  $\xi$ . The variable  $x_0$  has the probability distribution  $p_0(x_0)$ , whereas  $\xi$  is distributed according to a Gaussian distribution with mean zero and variance  $\sigma_{\xi}^2$ . The two parameters  $\sigma_{\text{eff}}^2$  and  $\sigma_{\xi}^2$  are given by the following set of self-consistency conditions

$$q = [\langle u \rangle^2]_{x_0, \xi}^{\text{av}}, \quad \Delta Q \equiv Q - q = [\langle (u - \langle u \rangle)^2 \rangle]_{x_0, \xi}^{\text{av}} \quad (29)$$

$$\sigma_{\text{eff}}^2 = \sigma^2 + \frac{\beta \Delta Q}{\alpha}, \quad \sigma_{\xi}^2 = \frac{q}{\alpha} \quad (30)$$

where the thermal averages  $\langle \dots \rangle$  over  $u$  are performed in the  $P_{\text{eff}}$  ensemble and the so-called quenched average  $[\dots]_{x_0, \xi}^{\text{av}}$  is over variables  $\xi$  and  $x_0$ .

In order to study the regularized least-squares reconstruction, we need to take the limits  $\beta \rightarrow \infty$ , and then

$\sigma \rightarrow 0$ . A nontrivial aspect of the zero temperature limit ( $\beta \rightarrow \infty$ ) is the quantity  $\beta \Delta Q$  in Eq. (30) that behaves differently in different phases of reconstruction. Using Eq. (29), this quantity is just  $\beta$  times the thermal fluctuation in  $u$ . The fluctuation-dissipation relation [23] implies that this quantity may be interpreted as a local susceptibility.

In our following considerations based on the zero temperature cavity method, we formally introduce a susceptibility and use its properties to give a more transparent derivation of the same equations via two step cavity method. We will outline the main points of the derivations in the next section, and refer the reader to Appendices A and B for the details.

#### IV. OUTLINE OF THE CAVITY APPROACH

The optimization problem associated with the regularized least-squared based reconstruction problem involves minimizing the energy function  $\mathcal{E}_0(\mathbf{x}) = \frac{(\mathbf{y} - \mathbf{H}\mathbf{x})^2}{2\sigma^2} + V(\mathbf{x})$ . For the noise free case, using  $\mathbf{y} = \mathbf{H}\mathbf{x}_0$ , the energy to be optimized may be rewritten as

$$\mathcal{E}(\mathbf{u}) = \frac{1}{2\sigma^2} \mathbf{u}^\top \mathbf{H}^\top \mathbf{H} \mathbf{u} + V(\mathbf{u} + \mathbf{x}_0). \quad (31)$$

where  $\mathbf{u} = \mathbf{x} - \mathbf{x}_0$ . Note that, unlike the function  $\mathcal{E}_0(\mathbf{x})$ , which is parametrized by known quantities (the data  $\mathbf{y}$  and the measurement matrix  $\mathbf{H}$ ) and can therefore be empirically optimized with respect to its argument, the closely related function  $\mathcal{E}(\mathbf{u}) = \mathcal{E}_0(\mathbf{u} + \mathbf{x}_0)$  depends on the knowledge of the original signal  $\mathbf{x}_0$ . The purpose of dealing with this function is *not* to provide an algorithm to estimate this signal given measured data, but to study the *statistical* behavior of this function and its minima over the distribution of problem instances, namely, input signals and the measurement matrices. For example, we can calculate the distribution of each component of the estimation error vector  $\mathbf{u}$ , given the distributions of  $\mathbf{x}_0$  and  $\mathbf{H}$ . We will be working with  $\mathcal{E}(u)$ , although the susceptibility for a particular problem instance, to be defined below, could be defined completely in terms of  $\mathcal{E}_0(\mathbf{x})$ .

In case this cost function reproduces the correct answer, the function  $\mathcal{E}(\mathbf{u})$  minimizes at  $\mathbf{u} = 0$ . Looking at the structure of  $\mathcal{E}(u)$  near zero tells us about potential shallow directions in parameter estimation error space, along which the cost function does rise significantly to reign in error. This failure could be quantified in terms of a susceptibility to error under linear perturbations to the cost function. The susceptibility provides a measure of robustness of the estimated parameters, and could indicate the trustworthiness of the reconstructed solution.

In addition, the cavity method derivation of the mean field equations involves considering the effect of altering a single variable or of imposing an additional data constraint. These modifications would be treated as ‘small’ perturbations to the large-scale optimization problem. The susceptibility, especially the local part of it, would

play a central role in computing the effect of these perturbations.

*Definition of Susceptibility:* As usual, let us consider a general regularization function  $V(\mathbf{x})$  for which there is a unique minimum to the cost function. Let the minimum of  $\mathcal{E}(\mathbf{u})$  be at  $\mathbf{u} = \hat{\mathbf{u}}$ . We introduce an augmented cost function

$$\mathcal{E}(\mathbf{u}; \mathbf{f}) = \frac{1}{2\sigma^2} \mathbf{u}^T \mathbf{H}^T \mathbf{H} \mathbf{u} + V(\mathbf{u} + \mathbf{x}_0) - \mathbf{f} \cdot \mathbf{u}. \quad (32)$$

with the variables  $\mathbf{f}$ , which are conjugate to  $\mathbf{u}$ . Optimizing  $\mathcal{E}(\mathbf{u}; \mathbf{f})$  will produce an  $\mathbf{f}$  dependent answer  $\mathbf{u} = \hat{\mathbf{u}}(\mathbf{f})$ . For small  $\mathbf{f}$  we expect

$$\hat{\mathbf{u}}(\mathbf{f}) = \hat{\mathbf{u}} + \chi \mathbf{f} + \dots \quad (33)$$

defining the susceptibility matrix  $\chi$ . Note that we want to take  $M, N \rightarrow \infty$  first before we take the  $\mathbf{f} \rightarrow 0$ , to define susceptibility. We also expect that

$$\min_{\mathbf{u}} \mathcal{E}(\mathbf{u}; \mathbf{f}) = \mathcal{E}(\hat{\mathbf{u}}) - \hat{\mathbf{u}}^T \mathbf{f} - \frac{1}{2} \mathbf{f}^T \chi \mathbf{f} + \dots \quad (34)$$

When  $\mathbf{H}$  is a large random matrix, we can make asymptotic estimates of the mean and the variance of different components of the susceptibility matrix  $\chi$ , following earlier work on singular values of random matrices [24, 25]. This is carried out in Appendix A. Note that only the diagonal terms  $\chi^{aa}$  have non-trivial means, whereas the off-diagonal terms average to zero. For diagonal terms, namely local susceptibilities, the average over all  $a$ 's,

$$\bar{\chi} \equiv \frac{1}{N} \sum_a \chi^{aa}, \quad (35)$$

is expected to be self averaging in the large  $M, N$  limit and be independent of the  $\mathbf{H}$  for a matrix chosen from the distribution  $\mathcal{P}(\mathbf{H})$ .

One should note,  $\chi^{ab}$ ,  $a \neq b$ , is an  $\mathbf{H}, \mathbf{x}_0$ -dependent number of the order  $1/\sqrt{M}$  for a particular choice of  $\mathbf{H}$  and of  $\mathbf{x}_0$ . It only vanishes after averaging over problem instances. Moreover, even if these  $O(1/\sqrt{M})$  terms are small compared to the  $O(1)$  diagonal terms, the off-diagonal terms have an important effect on the self-consistency equations via the so-called Onsager reaction term [26], as we will see in Appendix B.

*Removing a Variable Node:* For the ensuing discussion, it is useful to visualize the problem in terms of a bipartite graph (see Fig. 1), where the variables  $x_a$  are represented by circular nodes and the ‘constraints’ arising from each  $y_i$  (namely, the terms  $\frac{1}{2\sigma^2} (y_i - \sum_a H_{ia} x_a)^2 = \frac{1}{2\sigma^2} (\sum_a H_{ia} x_a)^2$  in the cost function) are represented by squares. Had we stuck to a finite temperature description, this graph would be the factor graph [27]. If we insist on satisfying the condition  $\mathbf{y} = \mathbf{H}\mathbf{x}$ , this graph could be thought of as a Tanner graph [28], with the circles being the variable nodes and the squares being the ‘check’ nodes. The system with  $N$  variables (circles) and  $M$  data

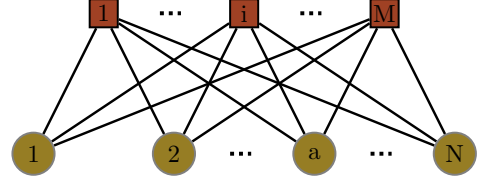


FIG. 1. Bipartite graph with variable nodes (circles) and constraint nodes (squares).

constraints (squares) would be represented as the  $(N, M)$  system. Our task is to relate properties of the  $(N, M)$  system to  $(N-1, M)$  system and obtain self-consistency conditions based on quantities that converge in the thermodynamic limit,  $N, M \rightarrow \infty$ .

We pick a particular node  $a$  and partition the cost function

$$\mathcal{E}(\mathbf{u}) = \frac{1}{2\sigma^2} \mathbf{u}^T \mathbf{H}^T \mathbf{H} \mathbf{u} + V(\mathbf{u} + \mathbf{x}_0) \quad (36)$$

into a contribution purely from the node, a term representing the interaction of the node variable with the rest of the system, and, lastly, the cost function of the  $(N-1, M)$  system:

$$\begin{aligned} \mathcal{E}(\mathbf{u}) = & \frac{1}{2\sigma^2} u_a^2 + U(u_a + x_{0a}) + \frac{1}{\sigma^2} u_a \mathbf{h}_a \cdot \sum_{b \setminus a} \mathbf{h}_b u_b \\ & + \frac{1}{2\sigma^2} \left( \sum_{b \setminus a} \mathbf{h}_b u_b \right)^2 + \sum_{b \setminus a} U(u_b + x_{0b}). \end{aligned} \quad (37)$$

Here, the  $a$ -th column of the  $\mathbf{H}$  matrix is being represented by the vector  $\mathbf{h}_a$ , and, the subscript  $\setminus a$  indicates that we leave out the node  $a$ . Moreover, we approximated  $\mathbf{h}_a^2$  by its average value

$$[\mathbf{h}_a^2]_{\mathbf{H}}^{\text{av}} = \sum_i [H_{ia}^2]_{\mathbf{H}}^{\text{av}} = \sum_{i=1}^M \frac{1}{M} = 1, \quad (38)$$

since  $\mathbf{h}_a^2$  is a sum of  $M$  terms and is self-averaging. The typical fluctuation of  $\mathbf{h}_a^2$  from its average value 1 asymptotically vanishes as  $O(1/\sqrt{M})$ .

The system without node ‘ $a$ ’, i.e., the system with a ‘cavity’ (see Fig. 2), will have its own optimum values  $u_b = \hat{u}_b$ , for all  $b \neq a$ . The variable  $u_a$  interacts with the rest of the system through the quantity  $\mathbf{h}_a \cdot \sum_{b \setminus a} \mathbf{h}_b u_b$ . The program of cavity method is to characterize the distribution of this quantity in terms of some parameters relating to the  $(N-1, M)$  system, and then use the fact that node ‘ $a$ ’ is statistically the same as every other node to relate these parameters to the distribution of  $u_a$ .

Since we are looking for the ground state, we minimize the expression in Eq. (37). This optimization becomes equivalent to the minimization of the following system with the node variable, the Onsager reaction term [26]

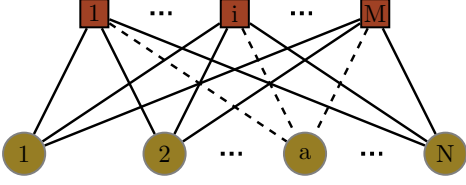


FIG. 2. The  $(N-1, M)$  cavity system. Node  $a$  has been removed from the system by removing the links to it.

and a contribution from the system with the ‘cavity’:

$$\min_{\mathbf{u}} \mathcal{E}(\mathbf{u}) = \min_{u_a} \left\{ \frac{1}{2\sigma^2} u_a^2 + U(u_a + x_{0a}) - \frac{1}{2\sigma^2} \frac{\bar{\chi}}{\alpha\sigma^2 + \bar{\chi}} u_a^2 + \frac{1}{\sigma^2} u_a \mathbf{h}_a \cdot \sum_{b \neq a} \mathbf{h}_b \hat{u}_b + \mathcal{E}_{\setminus a}(\hat{\mathbf{u}}_{\setminus a}) \right\}. \quad (39)$$

The detailed of this calculation is given in Appendix B. The Onsager term  $-\frac{1}{2\sigma^2} \frac{\bar{\chi}}{\alpha\sigma^2 + \bar{\chi}} u_a^2$  appears as a reaction term toward the variable  $u_a$  due to the adjustment of the other nodes after optimizing over them while holding  $u_a$  fixed. The coefficient of  $u_a^2$  turns out to be independent of  $a$  for large systems, because of self-averaging. We combine the  $u_a^2$  terms in Eq. (39) to get

$$\min_{\mathbf{u}} \mathcal{E}(\mathbf{u}) = \min_{u_a} \left\{ \frac{1}{2\sigma_{\text{eff}}^2} u_a^2 + U(u_a + x_{0a}) + \frac{1}{\sigma^2} u_a \mathbf{h}_a \cdot \sum_{b \neq a} \mathbf{h}_b \hat{u}_b + \mathcal{E}_{\setminus a}(\hat{\mathbf{u}}_{\setminus a}) \right\} \quad (40)$$

with  $\sigma_{\text{eff}}^2 = \sigma^2 + \bar{\chi}/\alpha$ .

If we did not have  $u_a$ , the system would be optimized at  $\hat{\mathbf{u}}_{\setminus a}$ . But now that  $u_a$  is coupled to  $\mathbf{h}_a \cdot \sum_{b \neq a} \mathbf{h}_b \hat{u}_b$ , one needs to characterize the distribution of this quantity. Let us define

$$\eta_a = -\mathbf{h}_a \cdot \sum_{b \neq a} \mathbf{h}_b \hat{u}_b = -\sum_{i=i}^M H_{ia} \hat{v}_i \quad (41)$$

where  $\hat{v}_i$  are the components of the residual vector  $\hat{\mathbf{v}} = \sum_{b \neq a} \mathbf{h}_b \hat{u}_b$ . Note that  $\mathbf{h}_a$  and  $\mathbf{v}$  are independent variables. Over the ensemble of problem instances,  $\eta_a$  has a distribution which is expected to be Gaussian, given that it is a sum of many contributions. All we need to do is to calculate the mean and the variance of this variable, using the independence of  $H_{ia}$ ’s and  $v_j$ ’s.

$$\begin{aligned} [\eta_a]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} &= -\sum_i [H_{ia}]_{\mathbf{H}}^{\text{av}} [\hat{v}_i]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} = 0 \\ [\eta_a^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} &= \sum_{i,j} [H_{ia} H_{ja}]_{\mathbf{H}}^{\text{av}} [\hat{v}_i \hat{v}_j]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} \\ &= \frac{1}{M} \sum_i [\hat{v}_i^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}}. \end{aligned} \quad (42)$$

Thus, all we need to know is the variance,  $[\hat{v}_i^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}}$ , of individual residuals for the  $(N-1, M)$  system.

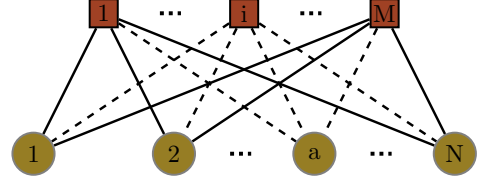


FIG. 3. The  $(N-1, M-1)$  cavity system. Node  $a$  and constraint  $i$  have been removed from the system by removing the links to them.

*Removing a Constraint Node:* An individual component  $\hat{v}_i$ , being a sum of many variables, is expected to be Gaussian over the problem instance ensemble. The mean turns out to be zero but the variance requires a more careful analysis. Because the variables  $H_{ib}$ ’s and  $\hat{u}_b$ ’s are strongly correlated,  $[\hat{v}_i^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} \neq \sum_{b \neq a} [H_{ib}^2]_{\mathbf{H}}^{\text{av}} [\hat{u}_b^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}}$ . To overcome this difficulty we need to replace  $\hat{u}_b$ ’s by variables that are independent of the  $i$ -th row of the  $\mathbf{H}$  matrix. A similar problem arises when using cavity method in the context of Hopfield neural networks [19].

In order to find these quantities, we go one step further by removing the constraint ‘ $i$ ’. The components of the error vector  $u'_b$  in the  $(N-1, M-1)$  system is indeed independent of the  $i$ -th row of the  $\mathbf{H}$  matrix. What remains to be done is to relate  $\hat{v}_i$  to  $u'_b$ ’s.

To find  $\hat{v}_i = \sum_{b \neq a} H_{ib} \hat{u}_b$ , we break up the minimization over  $\mathbf{u}_{\setminus a}$  into two steps:

$$\min_{\mathbf{u}_{\setminus a}} \mathcal{E}_{\setminus a}(\mathbf{u}_{\setminus a}) = \min_{v_i} \left\{ \min_{\substack{\mathbf{u}_{\setminus a} \\ \text{s.t. } \sum_{b \neq a} H_{ib} u_b = v_i}} \{ \mathcal{E}_{\setminus ai}(\mathbf{u}_{\setminus a}) \} + \frac{1}{2\sigma^2} v_i^2 \right\}. \quad (43)$$

In Appendix B we show that optimization of Eq. (43) is equivalent to

$$\min_{v_i} \left\{ \frac{\alpha}{2\bar{\chi}} (v_i - v'_i)^2 + \frac{1}{2\sigma^2} v_i^2 \right\}, \text{ with } v'_i = \sum_{b \neq a} H_{ib} u'_b. \quad (44)$$

We expect the first term to be minimized at  $v_i = v'_i$ , since minimization of the cost function  $\mathcal{E}_{\setminus ai}(\mathbf{u}_{\setminus a})$ , without the  $i$ -th constraint, would be at  $\mathbf{u}_{\setminus a} = \mathbf{u}'_{\setminus a}$ , making  $v_i = \sum_{b \neq a} H_{ib} u_b = \sum_{b \neq a} H_{ib} u'_b = v'_i$  at that point. The coefficient of  $(v_i - v'_i)^2$  depends only on  $\bar{\chi}$  and not on  $i$ , because of self-averaging once again. The second term forces the residual to be small, imposing correlations between  $H_{ib}$ ’s and  $u_b$ ’s. We can capture the effect of such correlations by considering how  $v_i$  is optimized when both terms are present.

Minimizing with respect to  $v_i$  gives us

$$\hat{v}_i = \frac{1}{1 + \frac{\bar{\chi}}{\alpha\sigma^2}} \sum_{b \neq a} H_{ib} u'_b. \quad (45)$$

The denominator  $(1 + \frac{\bar{\chi}}{\alpha\sigma^2})$  ‘scales down’ the unconstrained answer  $\sum_{b \neq a} H_{ib} u'_b$ . It is the same factor that relates  $\sigma^2$  to  $\sigma_{\text{eff}}^2$ .

Given that this result is true for any  $i$ 's, Eq. (39) becomes

$$\min_{\mathbf{u}} \mathcal{E}(\mathbf{u}) = \min_{u_a} \left\{ \frac{1}{2\sigma_{\text{eff}}^2} u_a^2 - \frac{1}{\sigma_{\text{eff}}^2} \xi_a u_a + U(u_a + x_{0a}) \right\} \quad (46)$$

with

$$\xi_a \equiv - \sum_i H_{ia} \sum_{b \neq a} H_{ib} u'_b \quad (47)$$

being a random Gaussian variable with mean zero and variance  $\sigma_{\xi}^2 \equiv [\xi_a^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} = \frac{q}{\alpha}$ . The quantity  $q \equiv \frac{1}{N-1} \sum_{b, c \neq a} [u'^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}}$  is the MSE for the  $(N-1, M-1)$  system. Insisting that  $q$  is also the MSE of the  $(N, M)$  system is one of the self-consistency conditions.

*Cavity Method Self-consistent Equations:* In summary, the zero temperature problem boils down to optimizing a collection of ‘independent’ variables

$$\hat{u}_a(f_a) = \underset{u_a}{\text{argmin}} \left\{ \frac{1}{2\sigma_{\text{eff}}^2} (u_a^2 - 2\xi_a u_a) + U(u_a + x_{0a}) - f_a u_a \right\} \quad (48)$$

using the same effective cost function as Eq. (28) in Sec. III, but with an additional linear perturbation. The variables  $\xi_a$  are chosen independently from  $\mathcal{N}(0, \sigma_{\xi}^2)$  and  $x_0$ 's are chosen independently from the probability distribution  $p_0(x_0)$ . With  $x_{0a}, \xi_a$  chosen randomly, we can obtain the distributions of  $u_a(0)$  and  $\chi^{aa} = \left. \frac{d\hat{u}_a(f_a)}{df_a} \right|_{f_a=0}$ .

The two parameters  $\sigma_{\text{eff}}^2$  and  $\sigma_{\xi}^2$  are decided by the following set of self-consistency conditions:

$$q = [u_a(0)^2]_{x_0, \xi}^{\text{av}}, \quad \bar{\chi} = [\chi^{aa}]_{x_0, \xi}^{\text{av}} \quad (49)$$

and

$$\sigma_{\text{eff}}^2 = \sigma^2 + \frac{\bar{\chi}}{\alpha}, \quad \sigma_{\xi}^2 = \frac{q}{\alpha}. \quad (50)$$

So far, we have analyzed the  $\sigma_{\xi} = 0$  case. The presence of additive noise can be handled easily by our method, with  $\hat{v}_i \equiv \sum_{b \neq a} H_{ib} \hat{u}_b + \zeta_i$  and  $v'_i \equiv \sum_{b \neq a} H_{ib} u'_b + \zeta_i$ . The effect of the additive noise could be absorbed in the  $\xi_a$  variable, with the new self-consistency condition for the variance  $\sigma_{\xi}^2$  would being:

$$\sigma_{\xi}^2 = \frac{q}{\alpha} + \sigma_{\zeta}^2. \quad (51)$$

In this formulation, we do not need to invoke temperature. The average local susceptibility  $\bar{\chi}$  plays the role of the quantity  $\beta \Delta Q$  in the replica approach. In our subsequent work [29], we will use  $\bar{\chi}$  to distinguish phases around zero-temperature critical point, including the Donoho-Tanner transition [8].

## V. MEAN FIELD THEORY AND FINITE SIZE $\ell_1$ -PENALTY RECONSTRUCTION

The mean field self-consistency equations derived above are for the thermodynamic limit, namely for

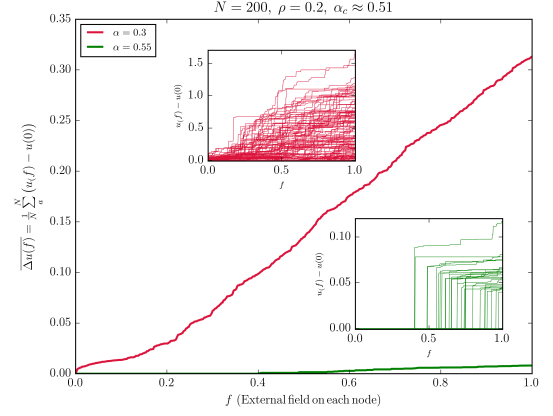


FIG. 4. Plot of the average error of the solution as a function of external field ‘ $f$ ’ on each node in two different regimes: green line being for perfect recovery regime and the red line corresponding error-prone one. The insets with corresponding colors are the responses of individual nodes, showing the staircase like behavior mentioned in the text.

$M, N \rightarrow \infty$ . Many of the quantities we define, like  $\bar{\chi}$  and  $\sigma_{\text{eff}}^2 (= \sigma^2 + \bar{\chi}/\alpha)$ , strictly make sense only after we take the problem size to infinity. How do these concepts arise in finite sized problems? To investigate this question, we look at the average local susceptibility  $\bar{\chi}$  for the important case of Basis Pursuit, which corresponds to  $V(\mathbf{x}) = \|\mathbf{x}\|_1$  and to  $\sigma \rightarrow 0$ .

In particular, we carry out the numerical experiment for minimization of  $\|\mathbf{u} + \mathbf{x}_0\|_1 - f u_a$ , for each  $a = 1, \dots, N$ , subject to  $\mathbf{H}\mathbf{u} = \mathbf{0}$ , by linear programming using the CVXOPT package [30]. We obtain the  $M \times N$  matrix  $\mathbf{H}$  by filling it with independent entries from a Gaussian distribution with mean zero and variance  $1/M$ . In this example, the size of the vector  $\mathbf{x}$  is  $N = 200$ , and contains  $K = 40$  randomly placed elements driven from a standard Gaussian distribution.

Since we solve a linear programming problem with the cost function perturbed by the linear term  $-f u_a$ , solutions are chosen from the extreme points of a convex polytope. As  $f$  is changed, the solution, would jump from one extreme point to another at particular thresholds. As a result, functions  $u_a(f)$  look like a set of staircases (see insets in Fig. 4).

To see the average local susceptibility emerge in the thermodynamic limit from these step functions, we need to compute the average response. The average parameter estimation error,  $\overline{\Delta u(f)} = \frac{1}{N} \sum_a (u_a(f) - u_a(0))$ , as a function of external field ‘ $f$ ’, shown for different regimes in Fig. 4, are strikingly different. In the high  $\alpha$  solution, the average error has no response to the external field up to a large threshold. However, in the low  $\alpha$  case, very small external fields can perturb the estimated solution to a new one. This is an indication of the robustness of the solution in the good recovery region and lack of it in



the error-prone regime, and hints at a phase transition in between. In this particular case, that transition takes place at  $\alpha = \alpha_c \approx 0.51$ .

To connect results shown in Fig. 4 to the average local susceptibility, we need to make linear fit to the average response  $\Delta u(f)$  as a function of  $f$  near  $f = 0$ . Note that differentiating  $u_a(f)$  with respect to  $f$  for the finite size system and then summing over  $a$  would give us a very different answer. In the good regime, the average local susceptibility, obtained by the fit, is approximately zero. As one decreases the number of constraints on the system past a threshold, this susceptibility becomes non-zero, as can be seen from the slope of the average response near zero, for the low  $\alpha$  solution in Fig. 4.

## VI. AFTERWORD

In this study, we directly treat the regularized least-squared optimization problem and show how to adapt the cavity method for doing mean field theory in the context. The mean field theory leads to a self-consistency condition on average mean squared error (MSE), since error in estimating one variable affects error in others. Careful derivation of the self-consistency condition, without using replica trick, involve accounting for subtle correlations in the system. To take care of these correlations, we needed a two-step cavity approach: one step removing a variable and then, another, removing a data constraint.

Although, we have emphasized the zero-temperature treatment, the cavity method can be used for finite temperature results as well. For completeness, we have provided the corresponding derivation in Appendix C. The key connection with the zero-temperature treatment is via the fluctuation-dissipation theorem [23], which relates thermal fluctuation with susceptibility.

The cavity approach looks at the behavior of the system for a particular choice of quenched variables,  $\mathbf{H}$  and  $\mathbf{x}_0$  in this case. In contrast, the replica approach centers on immediately averaging those quenched variables away. In the context of compressed sensing, one can imagine many problems, where the matrix  $\mathbf{H}$  is non-random. Currently there is no obvious way to extend the replica mean field treatment for such sensing matrices. The cavity method could be a more versatile tool in this regard. Extensions of this tool to other classes of compressed sensing problems would be a goal of future studies.

### Appendix A: Susceptibility and Conjugate Fields

We want to minimize the augmented cost function

$$\mathcal{E}(\mathbf{u}; \mathbf{f}) = \frac{1}{2\sigma^2} \mathbf{u}^T \mathbf{H}^T \mathbf{H} \mathbf{u} + V(\mathbf{u} + \mathbf{x}_0) - \mathbf{f} \cdot \mathbf{u}. \quad (\text{A1})$$

producing an  $\mathbf{f}$ -dependent minimum  $\mathbf{u} = \hat{\mathbf{u}}(\mathbf{f})$ . The susceptibility matrix  $\chi$  is defined by the small  $\mathbf{f}$  expansion:

$\hat{\mathbf{u}}(\mathbf{f}) = \hat{\mathbf{u}}(0) + \chi \mathbf{f} + \dots$ . If  $\mathcal{E}$  is differentiable, the optimum  $\hat{\mathbf{u}}(\mathbf{f})$  is the solution of

$$\mathbf{f} = \nabla_{\mathbf{u}} \mathcal{E}(\mathbf{u}) \quad (\text{A2})$$

Where  $\mathbf{f} = \mathbf{0}$ ,  $\mathbf{u}$  is at its optimal value  $\hat{\mathbf{u}}$ . Formally, if perturbation  $\mathbf{f}$  is small and  $\mathcal{E}$  is differentiable to higher orders, we can expect  $\delta \mathbf{u}$  to be small and, therefore, Taylor expand  $\mathcal{E}(\mathbf{u} + \delta \mathbf{u})$  around  $\mathbf{u} = \hat{\mathbf{u}}$

$$\mathcal{E}(\hat{\mathbf{u}} + \delta \mathbf{u}) = \mathcal{E}(\hat{\mathbf{u}}) + \frac{1}{2} \sum_{ab} \delta u_a \delta u_b \left. \frac{\partial^2 \mathcal{E}}{\partial u_a \partial u_b} \right|_{\mathbf{u}=\hat{\mathbf{u}}} + \dots \quad (\text{A3})$$

From (A2) and (A3), we can identify the inverse susceptibility  $(\chi^{-1})_{ab} = \left. \frac{\partial^2 \mathcal{E}}{\partial u_a \partial u_b} \right|_{\mathbf{u}=\hat{\mathbf{u}}}$  and can show that

$$\min_{\mathbf{u}} \mathcal{E}(\mathbf{u}; \mathbf{f}) = \mathcal{E}(\hat{\mathbf{u}}) - \hat{\mathbf{u}}^T \mathbf{f} - \frac{1}{2} \mathbf{f}^T \chi \mathbf{f} + \dots \quad (\text{A4})$$

It is simplest to study the properties of  $\chi$ , when the potential  $U(x)$  has continuous second derivatives. If we Taylor expand around the solution  $\mathbf{u} = \hat{\mathbf{u}}$ , we will have

$$\begin{aligned} \mathcal{E}(\hat{\mathbf{u}} + \delta \mathbf{u}; \mathbf{f}) = & \mathcal{E}(\hat{\mathbf{u}}; 0) + \frac{1}{2} \delta \mathbf{u}^T \left[ \frac{\mathbf{H}^T \mathbf{H}}{\sigma^2} + \mathbf{W}(\mathbf{x}) \right] \delta \mathbf{u} \\ & - \mathbf{f} \cdot (\hat{\mathbf{u}} + \delta \mathbf{u}) + \dots \end{aligned} \quad (\text{A5})$$

where  $W_{ab}(\mathbf{x}) = U''(\hat{u}_a + x_{0a}) \delta_{ab}$ . Optimizing over  $\delta \mathbf{u}$ , we see that the susceptibility matrix would be given by

$$\chi(\mathbf{x}, \mathbf{H}) = \left[ \frac{\mathbf{H}^T \mathbf{H}}{\sigma^2} + \mathbf{W}(\mathbf{x}) \right]^{-1}. \quad (\text{A6})$$

These statements are true, for fixed  $N$ , with  $\mathbf{f}$  and  $\delta \mathbf{u}$  going to zero. However, we are interested in the opposite limit,  $N \rightarrow \infty$  first and then taking  $\mathbf{f}, \delta \mathbf{u}$  small. We also need to deal with  $U(x)$  that is singular. The way we will treat this difficulty is as follows. We will keep  $U(x)$  to be smooth, take  $N \rightarrow \infty$  limit on  $\chi$  defined by Eq. A6, with the assumption that  $W_{aa}(\mathbf{x})$ 's come from a well-defined distribution. In that case, when  $\mathbf{H}$  is a large random matrix, we can make asymptotic estimates of the mean and the variance of different components of the susceptibility matrix  $\chi$ , using diagrammatic expansions and resummings.

To find the asymptotic behavior of  $\chi$ , we formally expand the RHS of Eq. (A6) in powers of  $\frac{\mathbf{H}^T \mathbf{H}}{\sigma^2}$  (see Fig. 5) and compute moments by averaging over  $H_{ia}$  diagrammatically. Namely, we expand

$$\begin{aligned} \chi = & \mathbf{W}^{-1} - \frac{1}{\sigma^2} \mathbf{W}^{-1} \mathbf{H}^T \mathbf{H} \mathbf{W}^{-1} \\ & + \frac{1}{\sigma^4} \mathbf{W}^{-1} \mathbf{H}^T \mathbf{H} \mathbf{W}^{-1} \mathbf{H}^T \mathbf{H} \mathbf{W}^{-1} - \dots \end{aligned} \quad (\text{A7})$$

and then compute moments of the form  $[\chi^{a_1 b_1} \chi^{a_2 b_2} \dots \chi^{a_k b_k} H_{i_1 c_1} \dots H_{i_l c_l}]_{\mathbf{H}}^{\text{av}}$  using Wick's theorem [31], since  $\mathbf{H}$  distribution is Gaussian with mean and covariance specified by Eq. (2) and Eq. (3), respectively. One can write  $[\chi(\mathbf{x}, \mathbf{H})]_{\mathbf{H}}^{\text{av}}$  as  $[\mathbf{W}(\mathbf{x}) - \Sigma(\mathbf{x}) \mathbf{I}_N]^{-1}$ ,

$$\chi_{aa} = a \frac{1}{W_{aa}^{-1}} a - a \frac{1}{\sigma^2} W_{aa}^{-1} H_{ai}^T H_{ia} W_{aa}^{-1} a + \dots$$

$$+ a \frac{1}{\sigma^4} W_{aa}^{-1} H_{ai}^T H_{ib} W_{bb}^{-1} H_{bj}^T H_{ja} W_{aa}^{-1} a + \dots$$

FIG. 5. The diagrammatic expansion of susceptibility.

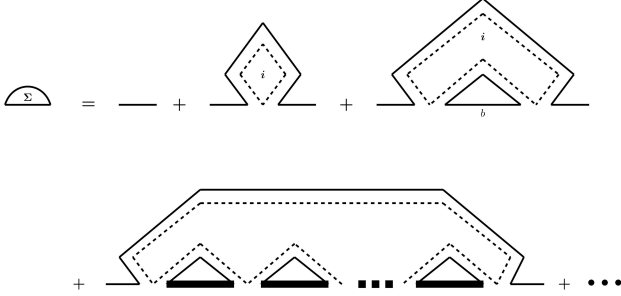


FIG. 6. Planar diagrams contributing to the self-energy.

where  $\Sigma(\mathbf{x})$  is a self-energy term. Using the fact that, in the large  $M, N$  limit, only the planar diagrams survive, the contributions to the self-energy are shown in Fig. 6 and can be re-summed as

$$\Sigma(\mathbf{x}) = -\frac{1}{\sigma^2} \frac{1}{1 + \frac{1}{M\sigma^2} \sum_a \chi^{aa}(\mathbf{x})} \quad (\text{A8})$$

Hence, the mean susceptibility (holding  $x_a$ 's fixed but averaging over  $\mathbf{H}$ ) is given by

$$\chi^{\text{av}}(\mathbf{x}) \equiv [\chi(\mathbf{x}, \mathbf{H})]_{\mathbf{H}}^{\text{av}} = \left[ \mathbf{W}(\mathbf{x}) + \frac{M}{M\sigma^2 + \text{Tr}[\chi^{\text{av}}(\mathbf{x})]} \mathbf{I}_N \right]^{-1}. \quad (\text{A9})$$

Covariance of  $\chi$  entries,  $[\chi^{ab}(\mathbf{x}, \mathbf{H}) \chi^{cd}(\mathbf{x}, \mathbf{H})]_{\mathbf{H}}^{\text{av}}$  could be computed using the diagrams in Fig. 7 and they are suppressed in the large  $M, N$  limit, since their contributions are  $O(\frac{1}{M}, \frac{1}{N})$ .

Therefore, we get the local susceptibilities, i.e. diagonal terms, to be:

$$[\chi^{aa}(\mathbf{x})]_{\mathbf{H}}^{\text{av}} = \left[ W_{aa}(x_a) + \frac{1}{\sigma^2 + \frac{\bar{\chi}(\mathbf{x})}{\alpha}} \right]^{-1} \quad (\text{A10})$$

$$\bar{\chi}(\mathbf{x}) \equiv \frac{1}{N} \sum_a [\chi^{aa}(\mathbf{x})]_{\mathbf{H}}^{\text{av}}. \quad (\text{A11})$$

Here,  $[\chi^{ab}]_{\mathbf{H}}^{\text{av}} = 0$  for  $a \neq b$ , but each  $\chi^{ab}$  is of order  $O(1/\sqrt{M})$ . In particular, we will need the correlation of  $\chi^{ab}$  with the corresponding matrix elements of  $\mathbf{H}^T \mathbf{H}$ .

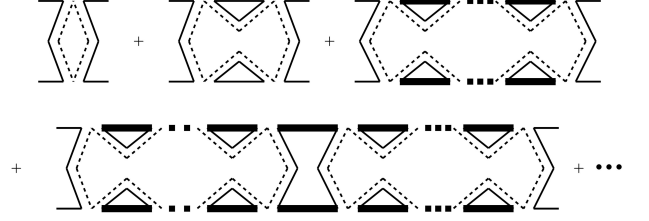


FIG. 7. The leading planar diagrams in covariance computation are of the order  $O(\frac{1}{M}, \frac{1}{N})$ , as can be seen from counting a factor of  $M$  or  $N$  for appropriate index loop, and counting a factor of  $\frac{1}{M}$  for each double-line contraction coming from averaging over the matrix elements.

Using the identity  $[\mathbf{W} + \frac{\mathbf{H}^T \mathbf{H}}{\sigma^2}] \chi = \mathbf{I}_N$ , we can prove a useful corollary of the result in Eq. (A9).

$$\begin{aligned} \frac{1}{\sigma^2} [\mathbf{H}^T \mathbf{H} \chi(\mathbf{x}, \mathbf{H})]_{\mathbf{H}}^{\text{av}} &= \mathbf{I}_N - \mathbf{W}(\mathbf{x}) \chi^{\text{av}}(\mathbf{x}) \\ &= \mathbf{I}_N - \mathbf{W}(\mathbf{x}) \left[ \mathbf{W}(\mathbf{x}) + \frac{M}{M\sigma^2 + \text{Tr}[\chi^{\text{av}}(\mathbf{x})]} \mathbf{I}_N \right]^{-1} \\ &= \frac{M}{M\sigma^2 + \text{Tr}[\chi^{\text{av}}(\mathbf{x})]} \left[ \mathbf{W} + \frac{M}{M\sigma^2 + \text{Tr}[\chi^{\text{av}}(\mathbf{x})]} \mathbf{I}_N \right]^{-1} \\ &= \frac{M \chi^{\text{av}}(\mathbf{x})}{M\sigma^2 + \text{Tr}[\chi^{\text{av}}(\mathbf{x})]} \\ &= \frac{\alpha \chi^{\text{av}}(\mathbf{x})}{\alpha \sigma^2 + \bar{\chi}(\mathbf{x})} \end{aligned} \quad (\text{A12})$$

In particular, Eq. (A12) implies

$$[\text{Tr}(\mathbf{H}^T \mathbf{H} \chi(\mathbf{x}, \mathbf{H}))]_{\mathbf{H}}^{\text{av}} = \frac{M\sigma^2 \bar{\chi}(\mathbf{x})}{\alpha \sigma^2 + \bar{\chi}(\mathbf{x})} \quad (\text{A13})$$

which we will use this identity in Appendix B.

Notice that many observations made here are independent of the assumption that  $U(x)$  has a continuous second derivative. For example, in the case of compressed sensing with  $U(x) = \lambda|x|$ , we could define a second-differentiable function  $U_\epsilon(x)$  such that  $\lim_{\epsilon \rightarrow 0} U_\epsilon(x) = U(x)$ , for example,  $U_\epsilon(x) = \sqrt{x^2 + \epsilon^2}$  or  $U_\epsilon(x) = \frac{1}{\epsilon} \log(2 \cosh(\epsilon x))$ . If  $x_a = x_{0a} + u_a$  goes to zero as  $\epsilon$  vanishes, then the corresponding  $W_{aa} = U''_\epsilon(x_a)$  diverges. However, the corresponding local susceptibility,  $\chi^{aa}$ , just becomes zero in this limit. Therefore, as  $\epsilon \rightarrow 0$ , the idea of using effective single variable optimization problems and determining the self-consistent distribution of  $x_a$  and  $\chi^{aa}$  remains meaningful. We just need to separate out the set of variables  $x_a$  for which  $W_{aa}$  diverges and treat this set carefully. As a consequence of  $\chi$  remaining well-defined in the  $\epsilon \rightarrow 0$  limit.

## Appendix B: Zero-temperature Cavity Method

### 1. Removing a Variable Node

We expand the following equation in terms of node ‘a’

$$\mathcal{E}(\mathbf{u}) = \frac{1}{2\sigma^2} \mathbf{u}^T \mathbf{H}^T \mathbf{H} \mathbf{u} + V(\mathbf{u} + \mathbf{x}_0) \quad (\text{B1})$$

which results in

$$\begin{aligned} \mathcal{E}(\mathbf{u}) = & \frac{1}{2\sigma^2} u_a^2 + U(u_a + x_{0a}) + \frac{1}{\sigma^2} u_a \mathbf{h}_a \cdot \sum_{b \setminus a} \mathbf{h}_b u_b \\ & + \frac{1}{2\sigma^2} \left( \sum_{b \setminus a} \mathbf{h}_b u_b \right)^2 + \sum_{b \setminus a} U(u_b + x_{0b}) \end{aligned} \quad (\text{B2})$$

The system with a cavity (at node ‘a’) has a new optimum values  $u_b = \hat{u}_b$ , for all  $b \neq a$ . In the complete system, the variable from node ‘a’, namely  $u_a$ , interacts with the rest of the variable via the term  $\mathbf{h}_a \cdot \sum_{b \setminus a} \mathbf{h}_b u_b$ . We rewrite Eq. (B2) representing the interaction between the node and the rest as a perturbation by a field:

$$\mathcal{E}(\mathbf{u}) = \frac{1}{2\sigma^2} u_a^2 + U(u_a + x_{0a}) + \mathcal{E}_{\setminus a}(\mathbf{u}_{\setminus a}) - \mathbf{u}_{\setminus a}^T \mathbf{f}_{\setminus a}. \quad (\text{B3})$$

The cost function of the  $(N-1, M)$  system is  $\mathcal{E}_{\setminus a}(\mathbf{u}_{\setminus a})$ . We identify  $(\mathbf{f}_{\setminus a})_b = -\frac{1}{\sigma^2} \mathbf{h}_b \cdot \mathbf{h}_a u_a$  to be the local force exerting on each node  $u_b$ , due to presence of node  $u_a$ . Since we are looking for the ground state, we minimize the expression in Eq. (B3)

$$\begin{aligned} \min_{\mathbf{u}} \mathcal{E}(\mathbf{u}) = & \min_{u_a, \mathbf{u}_{\setminus a}} \left\{ \frac{1}{2\sigma^2} u_a^2 + U(u_a + x_{0a}) \right. \\ & \left. + \mathcal{E}_{\setminus a}(\mathbf{u}_{\setminus a}) - \mathbf{u}_{\setminus a}^T \mathbf{f}_{\setminus a} \right\}. \end{aligned} \quad (\text{B4})$$

Given that  $\mathbf{h}_b \cdot \mathbf{h}_a$  is of the order  $1/\sqrt{M}$ ,  $\mathbf{f}_{\setminus a}$  is small, and we can invoke the definition of susceptibility  $\chi_{\setminus a}$  for the  $(N-1, M)$  system and use expansion of the minimized cost function (A4).

$$\begin{aligned} \min_{\mathbf{u}} \mathcal{E}(\mathbf{u}) = & \min_{u_a} \left\{ \frac{1}{2\sigma^2} u_a^2 + U(u_a + x_{0a}) \right. \\ & \left. + \mathcal{E}_{\setminus a}(\hat{\mathbf{u}}_{\setminus a}) - \hat{\mathbf{u}}_{\setminus a}^T \mathbf{f}_{\setminus a} - \frac{1}{2} \mathbf{f}_{\setminus a}^T \chi_{\setminus a} \mathbf{f}_{\setminus a} \right\} \end{aligned} \quad (\text{B5})$$

and plugging in  $(\mathbf{f}_{\setminus a})_b = -\frac{1}{\sigma^2} \mathbf{h}_b \cdot \mathbf{h}_a u_a$ , we get

$$\begin{aligned} \min_{\mathbf{u}} \mathcal{E}(\mathbf{u}) = & \min_{u_a} \left\{ \frac{1}{2\sigma_{\text{eff}}^2} u_a^2 + U(u_a + x_{0a}) \right. \\ & \left. + \frac{1}{\sigma^2} u_a \mathbf{h}_a \cdot \sum_{b \neq a} \mathbf{h}_b \hat{u}_b + \mathcal{E}_{\setminus a}(\hat{\mathbf{u}}_{\setminus a}) \right\} \end{aligned} \quad (\text{B6})$$

where

$$\frac{1}{\sigma_{\text{eff}}^2} = \frac{1}{\sigma^2} \left( 1 - \frac{1}{\sigma^2} \sum_{b, c \neq a} (\mathbf{h}_a \cdot \mathbf{h}_b)(\mathbf{h}_a \cdot \mathbf{h}_c) \chi_{\setminus a}^{bc} \right) \quad (\text{B7})$$

The quantity  $\sum_{b, c \neq a} (\mathbf{h}_b)_i (\mathbf{h}_c)_j \chi_{\setminus a}^{bc}$ , is independent of  $\mathbf{h}_a$ . As a result,  $\sum_{b, c \neq a} (\mathbf{h}_a \cdot \mathbf{h}_b)(\mathbf{h}_a \cdot \mathbf{h}_c) \chi_{\setminus a}^{bc}$  can be replaced by  $\sum_{b, c \neq a} (\mathbf{h}_b)_i (\mathbf{h}_c)_j \chi_{\setminus a}^{bc} / M$ , thanks to the self-averaging of  $(\mathbf{h}_a)_i (\mathbf{h}_a)_j$ . Using Eq. (A13) for the  $(N-1, M)$  system,

$$\sum_{b, c \neq a} \mathbf{h}_b \cdot \mathbf{h}_c \chi_{\setminus a}^{bc} = \frac{M \sigma^2 \bar{\chi}_{\setminus a}}{\alpha \sigma^2 + \bar{\chi}_{\setminus a}} \approx \frac{M \sigma^2 \bar{\chi}}{\alpha \sigma^2 + \bar{\chi}} \quad (\text{B8})$$

with the last step having to do with  $\bar{\chi}$  becoming independent of  $N, M$  asymptotically. Using this last relation Eq. (B8), in Eq. (B7) we get

$$\sigma_{\text{eff}}^2 = \sigma^2 + \frac{\bar{\chi}}{\alpha} \quad (\text{B9})$$

Looking at Eq. (B6), we still have to determine the size of the coupling of the node variable  $u_a$  to the rest of the system via  $\eta_a = -\mathbf{h}_a \cdot \hat{\mathbf{v}}$  with  $\hat{\mathbf{v}} = \sum_{b \neq a} \mathbf{h}_b \hat{u}_b$ . In Section IV, we showed that the first and the second moments of the  $\eta_a$  are:

$$\begin{aligned} [\eta_a]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} &= 0 \\ [\eta_a^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} &= \frac{1}{M} \sum_i [\hat{v}_i^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}}. \end{aligned} \quad (\text{B10})$$

The order  $k$  cumulants go as  $M^{1-k/2}$  and, for  $k > 2$ , they tend to zero as  $M$  goes to infinity. Therefore, we will stop with the variance and treat  $\eta_a$  as a zero-mean Gaussian variable. We still need the variance, for which we need a condition to determine  $[\hat{v}_i^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}}$ . This requires a second step of the cavity method.

### 2. Removing a Constraint Node

The subtlety in determining  $[\hat{v}_i^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}}$  involves accounting for correlation between matrix elements  $H_{ib}$  and the optimal values  $\hat{u}_b$  of the  $(N-1, M)$  system. To do this, we need to set up an  $(N-1, M-1)$  system with the constraint ‘i’ removed (see Fig. 3). To find  $\hat{v}_i = \sum_{b \neq a} H_{ib} \hat{u}_b$ , we write the minimization over  $\mathbf{u}_{\setminus a}$  as follows:

$$\min_{\mathbf{u}_{\setminus a}} \mathcal{E}_{\setminus a}(\mathbf{u}_{\setminus a}) = \min_{v_i} \left\{ \min_{\substack{\mathbf{u}_{\setminus a} \\ \text{s.t. } \sum_{b \neq a} H_{ib} u_b = v_i}} \{ \mathcal{E}_{\setminus a}(\mathbf{u}_{\setminus a}) \} + \frac{1}{2\sigma^2} v_i^2 \right\} \quad (\text{B11})$$

with the first minimization being a constrained one for the  $(N-1, M-1)$  system, subject to  $\sum_{b \neq a} H_{ib} u_b = v_i$ , and the second one being over  $v_i$ . The cost function for the system without nodes  $a, i$  is represented by  $\mathcal{E}(\mathbf{u}_{\setminus a})_{\setminus i}$ . The term  $\frac{1}{2\sigma^2} v_i^2$  represents the constraint coming from the  $i$ -th observation. Had we done an unconstrained optimization of  $\mathcal{E}_{\setminus ai}(\mathbf{u}_{\setminus a})$ , the optimum  $\hat{\mathbf{u}}_{\setminus a}$  would be independent of  $H_{ib}$ . Trying to keep  $v_i$  small perturbs this solution by a small amount and induces correlation with  $H_{ib}$ . Our strategy would be to compute the effect of perturbation in terms of the system susceptibility.

In order to do constrained minimization, we use the Lagrange multiplier method

$$\min_{\mathbf{u}_{\setminus a}} \mathcal{E}_{\setminus a}(\mathbf{u}_{\setminus a}) = \max_{\gamma_i} \min_{\mathbf{u}_{\setminus a}, v_i} \{ \mathcal{E}_{\setminus ai}(\mathbf{u}_{\setminus a}) + \frac{1}{2\sigma^2} v_i^2 - \gamma_i (v_i - \sum_{b \neq a} H_{ib} u_b) \}. \quad (\text{B12})$$

Minimizing Eq. (B12) with respect to  $v_i$  we get  $v_i = \sigma^2 \gamma_i$ , and making that substitution for  $v_i$  into the cost function we get

$$\min_{\mathbf{u}_{\setminus a}} \mathcal{E}_{\setminus a}(\mathbf{u}_{\setminus a}) = \max_{\gamma_i} \min_{\mathbf{u}_{\setminus a}} \{ \mathcal{E}_{\setminus a}(\mathbf{u}_{\setminus a}) - \frac{1}{2} \sigma^2 \gamma_i^2 - \mathbf{u}_{\setminus a}^T \mathbf{g} \} \quad (\text{B13})$$

$$= \max_{\gamma_i} \{ -\frac{1}{2} \sigma^2 \gamma_i^2 + \mathcal{E}_{\setminus i}^*(\mathbf{g}) \} \quad (\text{B14})$$

with  $g_b = -\gamma_i H_{ib}$  and with  $\mathcal{E}_{\setminus i}^*(\mathbf{g})$  is defined as

$$\mathcal{E}_{\setminus i}^*(\mathbf{g}) = \min_{\mathbf{u}_{\setminus a}} \{ \mathcal{E}_{\setminus ai}(\mathbf{u}_{\setminus a}) - \mathbf{u}_{\setminus a}^T \mathbf{g} \} \quad (\text{B15})$$

where the presence of  $\mathbf{g}$  alters the optimal  $\mathbf{u}_{\setminus a}$  from the unconstrained optimum  $\mathbf{u}'_{\setminus a}$ . Since each component of  $\mathbf{g}$  is small ( $O(1/\sqrt{M})$ ), we can expand around  $\mathbf{u}'_{\setminus a}$  using  $\chi_{\setminus ai}$ , the susceptibility of the  $(N-1, M-1)$  system, as in Eq. (A4). Therefore,  $\mathcal{E}_{\setminus i}^*(\mathbf{g})$  can be written as

$$\mathcal{E}_{\setminus i}^*(\mathbf{g}) = \mathcal{E}_{\setminus i}(\mathbf{u}'_{\setminus a}) - \mathbf{u}'_{\setminus a}^T \mathbf{g} - \frac{1}{2} \mathbf{g}^T \chi_{\setminus ai} \mathbf{g} + \dots \quad (\text{B16})$$

Now, Eq. (B14) becomes

$$\min_{\mathbf{u}_{\setminus a}} \mathcal{E}_{\setminus a}(\mathbf{u}_{\setminus a}) = \min_{\gamma_i} \{ -\frac{1}{2} \sigma^2 \gamma_i^2 + \mathcal{E}_{\setminus ai}(\mathbf{u}'_{\setminus a}) - \mathbf{u}'_{\setminus a}^T \mathbf{g} - \frac{1}{2} \mathbf{g}^T \chi_{\setminus ai} \mathbf{g} \} \quad (\text{B17})$$

The quadratic term  $\mathbf{g}^T \chi_{\setminus ai} \mathbf{g} = \gamma_i^2 \sum_{ij} H_{ib} H_{ic} \chi_{\setminus ai}^{bc}$  can be simplified because of self-averaging. We have

$$\sum_{ij} [H_{ib} H_{ic}]_{\mathbf{H}}^{\text{av}} \chi_{\setminus ai}^{bc} = \frac{1}{M} \sum_b \chi_{\setminus ai}^{bb} \approx \frac{\bar{\chi}}{\alpha}, \quad (\text{B18})$$

once more using the fact that the average local susceptibility  $\bar{\chi}$  is nearly the same for the  $(N, M)$  system and the  $(N-1, M-1)$  system.

Putting everything together

$$\min_{\mathbf{u}_{\setminus a}} \mathcal{E}_{\setminus a}(\mathbf{u}_{\setminus a}) = \max_{\gamma_i} \{ -\frac{\sigma^2}{2} (1 + \frac{\bar{\chi}}{\alpha \sigma^2}) \gamma_i^2 + \gamma_i \sum_{b \neq a} H_{ib} u'_b \}, \quad (\text{B19})$$

maximizing with respect to  $\gamma_i$  and then using  $v_i = \sigma^2 \gamma_i$  gives us

$$v_i = \frac{1}{1 + \frac{\bar{\chi}}{\alpha \sigma^2}} \sum_{b \neq a} H_{ib} u'_b. \quad (\text{B20})$$

Since this result is true for any  $i$ 's, Eq. (B6) becomes

$$\min_{\mathbf{u}} \mathcal{E}(\mathbf{u}) = \min_{u_a} \{ \frac{1}{2\sigma_{\text{eff}}^2} u_a^2 - \frac{\xi}{\sigma^2 (1 + \frac{\bar{\chi}}{\alpha \sigma^2})} u_a + U(u_a + x_{0a}) \} \quad (\text{B21})$$

with

$$\xi \equiv - \sum_i H_{ia} \sum_{b \neq a} H_{ib} u'_b \quad (\text{B22})$$

being a random Gaussian variable with mean zero and variance

$$\begin{aligned} \sigma_{\xi}^2 &\equiv [\xi^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} = \sum_{i,j} [H_{ia} H_{ja}]_{\mathbf{H}}^{\text{av}} \sum_{b,c \neq a} [H_{ib} H_{jc}]_{\mathbf{H}}^{\text{av}} [u'_b u'_c]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} \\ &= \sum_{i,j} \frac{\delta_{ij}}{M} \sum_{b,c \neq a} \frac{\delta_{ij} \delta_{bc}}{M} [u'_b u'_c]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} \\ &= \frac{1}{M} \sum_{b \neq a} [u_b'^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}} = \frac{q}{\alpha} \end{aligned} \quad (\text{B23})$$

thanks to  $H_{ja}$  and  $H_{ib}$  being independent for  $a \neq b$ , as well as  $u'_b$ 's being independent of those matrix elements. Here the quantity  $q \equiv \frac{1}{N-1} \sum_{b,c \neq a} [u'^2]_{\mathbf{x}_0, \mathbf{H}}^{\text{av}}$  is the MSE for

the  $(N-1, M-1)$  system. Taking into consider that  $q$  is a self-averaging quantity and is asymptotically the same for the  $(N, M)$  system, we get the independent single variable optimization

$$\min_{u_a} \{ \frac{1}{2\sigma_{\text{eff}}^2} (u_a^2 - 2\xi_a u_a) + U(u_a + x_{0a}) \} \quad (\text{B24})$$

with  $\xi_a \in \mathcal{N}(0, \sigma_{\xi}^2)$  and  $\sigma_{\text{eff}}^2 = \sigma^2 + \bar{\chi}/\alpha$ .

### Appendix C: Finite Temperature Cavity Method

In this section, we solve the finite temperature problem formulated in Sec. II via the cavity method. With the cost function written in terms of  $\mathbf{u}$  as

$$\mathcal{E}(\mathbf{u}) = \frac{1}{2\sigma^2} (\mathbf{H}\mathbf{u})^2 + V(\mathbf{u} + \mathbf{x}_0) \quad (\text{C1})$$

we define the Boltzmann distribution  $P(\mathbf{u}|\mathbf{H}, \mathbf{x}_0)$ :

$$P(\mathbf{u}|\mathbf{H}, \mathbf{x}_0) = \frac{1}{Z(\beta|\mathbf{H}, \mathbf{x}_0)} e^{-\beta \mathcal{E}} \quad (\text{C2})$$

with the normalization factor/partition function given by

$$Z(\beta|\mathbf{H}, \mathbf{x}_0) = \int d\mathbf{u} e^{-\beta \mathcal{E}} \quad (\text{C3})$$

We now apply the first step of the two-step cavity method. First, we rewrite  $\mathcal{E}$  as an interaction between variable  $u_a$  and the rest of the variables

$$\mathcal{E}(\mathbf{u}) = \frac{1}{2\sigma^2} \mathbf{h}_a^2 u_a^2 + \frac{1}{\sigma^2} u_a \mathbf{h}_a \cdot \sum_{b \neq a} \mathbf{h}_b u_b + U(u_a + x_{0a}) + \mathcal{E}_{\setminus a}(\mathbf{u}_{\setminus a}) \quad (\text{C4})$$

By defining

$$\eta_a \equiv - \frac{\mathbf{h}_a \cdot \sum_{b \neq a} \mathbf{h}_b u_b}{\mathbf{h}_a^2} \quad (\text{C5})$$

and using  $\mathbf{h}_a^2 = 1 + O(\frac{1}{\sqrt{M}})$  we have

$$\mathcal{E} = \frac{1}{2\sigma^2}(u_a^2 - 2u_a\eta_a) + U(u_a + x_{0a}) + \mathcal{E}_{\setminus a}(\mathbf{u}_{\setminus a}) \quad (\text{C6})$$

with  $\mathbf{u}_{\setminus a}$ ,  $\mathcal{E}_{\setminus a}$  etc defined as in Sec. IV. Equation (C6) indicates that the variable  $u_a$  interacts with all the others only through  $\eta_a$ . Therefore, we rewrite the marginal distribution  $P(u_a)$  as an integral over the joint distribution of  $\eta_a$  and  $u_a$ ,  $P(u_a, \eta_a)$ .

$$P(u_a) = \frac{1}{Z} \int d\mathbf{u}_{\setminus a} e^{-\beta\mathcal{E}} = \int d\eta_a P(u_a, \eta_a) \quad (\text{C7})$$

where

$$P(u_a, \eta_a) = \frac{1}{Z} \int d\mathbf{u}_{\setminus a} \delta(\eta_a + \mathbf{h}_a \cdot \sum_{b \neq a} \mathbf{h}_b u_b) e^{-\beta\mathcal{E}} \quad (\text{C8})$$

for all  $a = 1, \dots, N$ . Now we introduce a cavity ‘field’ distribution of  $\eta_a$  at the removed node  $a$  as

$$P_{\setminus a}(\eta_a) = \frac{1}{Z_{\setminus a}} \int d\mathbf{u} \delta(\eta_a + \mathbf{h}_a \cdot \sum_{b \neq a} \mathbf{h}_b u_b) e^{-\beta\mathcal{E}_{\setminus a}}. \quad (\text{C9})$$

By comparing (C8) and (C9), we get

$$P(u_a) = \frac{\int d\eta_a \exp\left[-\beta\left\{\frac{(u_a^2 - 2u_a\eta_a)}{2\sigma^2} + U(u_a + x_{0a})\right\}\right] P_{\setminus a}(\eta_a)}{\int du_a d\eta_a \exp\left[-\beta\left\{\frac{(u_a^2 - 2u_a\eta_a)}{2\sigma^2} + U(u_a + x_{0a})\right\}\right] P_{\setminus a}(\eta_a)} \quad (\text{C10})$$

The assumption of continuity of the global ground state, even in the presence of the cavity after removing node  $a$ , is equivalent to the replica symmetric hypothesis. This is a valid assumption when the penalty function  $V$  is convex. Therefore, in the limit of  $N \rightarrow \infty$ , even if the nodes of  $(N-1, M)$  system are weakly correlated,  $\eta_a$  is still a sum of many variables and  $P(\eta_a)_{\setminus a}$  can well be approximated by a Gaussian distribution.

$$P_{\setminus a}(\eta_a) \propto e^{-\frac{(\eta_a - \langle \eta_a \rangle_{\setminus a})^2}{2\langle \delta \eta_a^2 \rangle_{\setminus a}}} \quad (\text{C11})$$

Then (C10) becomes

$$P(u_a) = \frac{\exp\left\{-\frac{\beta}{2\sigma^2}\left(1 - \frac{\beta}{\sigma^2}\langle \delta \eta_a^2 \rangle_{\setminus a}\right)u_a^2 + \frac{\beta u_a}{\sigma^2}\langle \eta_a \rangle_{\setminus a} - \beta U(u_a + x_{0a})\right\}}{\int du_a \exp\left\{-\frac{\beta}{2\sigma^2}\left(1 - \frac{\beta}{\sigma^2}\langle \delta \eta_a^2 \rangle_{\setminus a}\right)u_a^2 + \frac{\beta u_a}{\sigma^2}\langle \eta_a \rangle_{\setminus a} - \beta U(u_a + x_{0a})\right\}} \quad (\text{C12})$$

Therefore, only the thermal averages  $\langle \eta_a \rangle_{\setminus a}$  and the thermal fluctuation strength  $\langle \delta \eta_a^2 \rangle_{\setminus a} = \langle (\eta_a - \langle \eta_a \rangle_{\setminus a})^2 \rangle_{\setminus a}$  of the field  $\eta_a$  for the distribution  $P_{\setminus a}(\eta_a)$  are left to be computed. In that process the effects of (weak) correlation between the  $u_a$ ’s have to be accounted for. Define, as in Sec. IV,

$$v_i = \sum_{b \neq a} H_{ib} u_b \quad (\text{C13})$$

and utilize our definition,

$$\eta_a = - \sum_i H_{ia} v_i \quad (\text{C14})$$

then we arrive at

$$\langle \eta_a \rangle_{\setminus a} = - \sum_i H_{ia} \langle v_i \rangle \quad (\text{C15})$$

and

$$\begin{aligned} \langle \delta \eta_a^2 \rangle_{\setminus a} &= \sum_{ij} H_{ia} H_{ja} \langle \delta v_i \delta v_j \rangle \\ &\approx \sum_{ij} \frac{1}{M} \delta_{ij} \langle \delta v_i \delta v_j \rangle = \frac{1}{M} \sum_i \langle \delta v_i^2 \rangle \end{aligned} \quad (\text{C16})$$

Having done that we need to compute  $\langle v_i \rangle$  and  $\langle \delta v_i^2 \rangle$ . To do so, this time in addition to site  $a$  we exclude site  $i$ . Hence from (C6) we get

$$\mathcal{E}_{\setminus a}(\mathbf{u}_{\setminus a}) = \frac{1}{2\sigma^2} v_i^2 + \mathcal{E}_{\setminus ai}(\mathbf{u}_{\setminus a}) \quad (\text{C17})$$

After carrying out the same computation as in (C7), (C8), and (C10) for the marginal distribution  $Q_{\setminus a}(v_i)$ , we arrive at

$$Q_{\setminus a}(v_i) = \frac{\exp\left\{-\frac{\beta}{2\sigma^2} v_i^2 - \frac{(v_i - \langle v_i \rangle_{\setminus ai})^2}{2\langle \delta v_i^2 \rangle_{\setminus ai}}\right\}}{\int dv_i \exp\left\{-\frac{\beta}{2\sigma^2} v_i^2 - \frac{(v_i - \langle v_i \rangle_{\setminus ai})^2}{2\langle \delta v_i^2 \rangle_{\setminus ai}}\right\}} \quad (\text{C18})$$

Therefore

$$Q_{\setminus a}(v_i) = \frac{\exp\left\{-\frac{\beta}{2\sigma^2}\left(1 + \frac{\sigma^2}{\beta\langle \delta v_i^2 \rangle_{\setminus ai}}\right)\left(v_i - \frac{\langle v_i \rangle_{\setminus ai}}{1 + \frac{\beta\langle \delta v_i^2 \rangle_{\setminus ai}}{\sigma^2}}\right)^2\right\}}{\int dv_i \exp\left\{-\frac{\beta}{2\sigma^2}\left(1 + \frac{\sigma^2}{\beta\langle \delta v_i^2 \rangle_{\setminus ai}}\right)\left(v_i - \frac{\langle v_i \rangle_{\setminus ai}}{1 + \frac{\beta\langle \delta v_i^2 \rangle_{\setminus ai}}{\sigma^2}}\right)^2\right\}} \quad (\text{C19})$$

and then  $\langle \delta v_i^2 \rangle$  is

$$\langle \delta v_i^2 \rangle = \frac{1}{\frac{\beta}{\sigma^2}\left(1 + \frac{\sigma^2}{\beta\langle \delta v_i^2 \rangle_{\setminus ai}}\right)} = \frac{\langle \delta v_i^2 \rangle_{\setminus ai}}{1 + \frac{\beta\langle \delta v_i^2 \rangle_{\setminus ai}}{\sigma^2}} \quad (\text{C20})$$

and  $\langle v_i \rangle$  is at

$$\langle v_i \rangle = \frac{\langle v_i \rangle_{\setminus ai}}{1 + \frac{\beta\langle \delta v_i^2 \rangle_{\setminus ai}}{\sigma^2}}. \quad (\text{C21})$$

Notice how both these moments for the  $(N-1, M)$  system is scaled down by the same factor, when compared to the moments for the  $(N-1, M-1)$  system. Using arguments similar to the fluctuation-dissipation [23] theorem, we could show that the change in  $\langle v_i \rangle$  due a change in  $\langle v_i \rangle_{\setminus ai}$ , susceptibility of sorts, is closely related to  $\langle \delta v_i^2 \rangle_{\setminus ai}^{-2}$  times  $\langle \delta v_i^2 \rangle$ , with the first term of the product playing the role of temperature.

Carrying on, we get

$$\begin{aligned} \langle \delta v_i^2 \rangle_{\setminus ai} &= \sum_{b, c \neq a} H_{ib} H_{ic} \langle \delta u_b \delta u_c \rangle_{\setminus ai} \\ &= \sum_{b, c \neq a} \frac{1}{M} \delta_{bc} \langle \delta u_b \delta u_c \rangle_{\setminus ai} + O\left(\frac{N}{M^{3/2}}, \frac{N^{1/2}}{M}\right) \\ &\approx \frac{1}{M} \sum_{b \neq a} \langle \delta u_b^2 \rangle_{\setminus ai} \end{aligned} \quad (\text{C22})$$

since the  $(N-1, M-1)$  system, indicated by the subscript ' $\backslash ai$ ', is independent of  $H_{ib}$  and  $H_{ic}$ ,  $H_{ib}H_{ic} = \frac{1}{M}\delta_{bc} + O(\frac{1}{M})$  fluctuations, and  $\langle \delta u_b \delta u_c \rangle_{\backslash ai} \sim O(\frac{1}{\sqrt{M}}, \frac{1}{\sqrt{N}})$  when  $b \neq c$ , indicating that nodes are only weakly correlated.

To make connection with the notation in Sec. III, let us introduce  $\Delta Q$

$$\Delta Q \equiv \frac{1}{N} \sum_a \langle \delta u_a^2 \rangle \approx \frac{1}{N-1} \sum_{b \neq a} \langle \delta u_b^2 \rangle_{\backslash ai}, \quad (\text{C23})$$

the second approximate equality becoming exact in the thermodynamic limit. Then, we have

$$\langle \delta v_i^2 \rangle_{\backslash ai} = \Delta Q / \alpha. \quad (\text{C24})$$

Therefore from (C16), (C20), and (C23)

$$\frac{\beta}{\sigma^2} \langle \delta \eta_a^2 \rangle_{\backslash a} = \frac{1}{(1 + \frac{\sigma^2}{\beta \Delta Q / \alpha})} \quad (\text{C25})$$

and from (C15), (C21), and (C23)

$$\langle \eta_a \rangle_{\backslash a} = \frac{\sum_i H_{ia} \sum_{b \neq a} H_{ib} \langle u_b \rangle_{\backslash ai}}{(1 + \frac{\beta \Delta Q}{\alpha \sigma^2})}. \quad (\text{C26})$$

Moreover, we define

$$\xi_a \equiv \sum_i H_{ia} \sum_{b \neq a} H_{ib} \langle u_b \rangle_{\backslash ai} \quad (\text{C27})$$

which has variance  $\sigma_\xi^2 = q/\alpha$  with  $q$

$$q = \frac{1}{N} \sum_a \langle u_a \rangle^2 \approx \frac{1}{N-1} \sum_b \langle u_b \rangle_{\backslash ai}^2 \quad (\text{C28})$$

being the mean squared error. Therefore, by plugging (C26) and (C27) into Eq. (C12), the marginal distribution for single variable  $u_a$  becomes

$$P(u_a) = \frac{\exp\{-\frac{\beta}{2\sigma_{\text{eff}}^2}(u_a^2 - 2u_a\xi_a) - \beta U(x_{0a} + u_a)\}}{\int du_a \exp\{-\frac{\beta}{2\sigma_{\text{eff}}^2}(u_a^2 - 2u_a\xi_a) - \beta U(x_{0a} + u_a)\}} \quad (\text{C29})$$

with  $\sigma_{\text{eff}}^2 = \sigma^2(1 + \frac{\beta \Delta Q}{\alpha \sigma^2})$ , and the effective cost function for the individual node is

$$\mathcal{E}(u_a) = \frac{1}{2\sigma_{\text{eff}}^2}(u_a^2 - 2u_a\xi_a) + U(x_{0a} + u_a) \quad (\text{C30})$$

Therefore, with  $\mathcal{E}$  replaced by a set of effectively decoupled nodes, and the sum over index  $a$  replaced by a quenched average over  $\xi_a, x_{0a}$ . As a result, the self-consistency conditions for the MSE

$$q = \frac{1}{N} \sum_{a=1}^N \langle u_a \rangle^2 \quad (\text{C31})$$

and for

$$\Delta Q = \frac{1}{N} \sum_{a=1}^N \langle \delta u_a^2 \rangle \quad (\text{C32})$$

reduce to

$$q = [\langle u \rangle_{\text{eff}}^2]_{\xi, x_0}^{\text{av}} \quad (\text{C33})$$

and

$$\Delta Q = [\langle \delta u^2 \rangle_{\text{eff}}]_{\xi, x_0}^{\text{av}} \quad (\text{C34})$$

where the thermal average  $\langle \dots \rangle_{\text{eff}}$  is performed with respect to the effective individual node distribution (27) and  $[\dots]_{\xi, x_0}^{\text{av}}$  is the quenched average over variables  $\xi, x_0$ , with  $\xi$  drawn from  $\mathcal{N}(0, q/\alpha)$  and signal  $x_0$  drawn independently from a distribution  $P(x_0)$ . These self-consistency equations are exactly the same those from the replica symmetric ansatz in Sec. III.

## ACKNOWLEDGMENTS

M.R. and A.M.S. acknowledge the hospitality of the Simons Center for Data Analysis. A.M.S. thanks Pankaj Mehta for comments on an earlier version of the manuscript. This work was supported by the National Science Foundation INSPIRE (track 1) award 1344069 to P.P.M. and A.M.S.

- 
- [1] S. Weisberg, *Applied Linear Regression*, vol. 528 (John Wiley & Sons, 2005).
  - [2] R. Tibshirani, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288 (1996).
  - [3] A. E. Hoerl, *Chemical Engineering Progress* **58**, 54 (1962).
  - [4] A. N. Tikhonov, *Dokl. Akad. Nauk SSSR* **39**, 195 (1943).
  - [5] S. S. Chen, D. L. Donoho, and M. A. Saunders, *SIAM*

- Journal on Scientific Computing* **20**, 33 (1998).
- [6] E. J. Candès, J. Romberg, and T. Tao, *Information Theory, IEEE Transactions on* **52**, 489 (2006).
- [7] D. L. Donoho, *Information Theory, IEEE Transactions on* **52**, 1289 (2006).
- [8] D. L. Donoho and J. Tanner, *Proceedings of the National Academy of Sciences of the United States of America* **102**, 9446 (2005).

- [9] D. L. Donoho and J. Tanner, in *Information Sciences and Systems, 2006 40th Annual Conference on* (IEEE, 2006), pp. 202–206.
- [10] S. Rangan, A. K. Fletcher, and V. K. Goyal, *Information Theory, IEEE Transactions on* **58**, 1902 (2012).
- [11] D. Guo and S. Verdú, *Information Theory, IEEE Transactions on* **51**, 1983 (2005).
- [12] T. Tanaka, *Information Theory, IEEE Transactions on* **48**, 2888 (2002).
- [13] Y. Kabashima, T. Wadayama, and T. Tanaka, *Journal of Statistical Mechanics: Theory and Experiment* **2009**, L09003 (2009).
- [14] S. Ganguli and H. Sompolinsky, *Physical review letters* **104**, 188701 (2010).
- [15] M. Mézard, G. Parisi, and M. Virasoro, *Europhys. Lett* **1**, 77 (1986).
- [16] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond*, vol. 9 (World scientific Singapore, 1987).
- [17] R. Mulet, A. Pagnani, M. Weigt, and R. Zecchina, *Physical Review Letters* **89**, 268701 (2002).
- [18] M. Mézard, G. Parisi, and R. Zecchina, *Science* **297**, 812 (2002).
- [19] M. Shamir and H. Sompolinsky, *Physical Review E* **61**, 1839 (2000).
- [20] A. Braunstein, M. Mézard, and R. Zecchina, *Random Structures & Algorithms* **27**, 201 (2005).
- [21] M. Mezard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, 2009).
- [22] D. L. Donoho, A. Maleki, and A. Montanari, *Proceedings of the National Academy of Sciences* **106**, 18914 (2009).
- [23] R. Kubo, *Reports on Progress in Physics* **29**, 255 (1966).
- [24] A. Sengupta and P. P. Mitra, *Physical Review E* **60**, 3389 (1999).
- [25] A. M. Sengupta and P. P. Mitra, *Journal of Statistical Physics* **125**, 1223 (2006).
- [26] L. Onsager, *Journal of the American Chemical Society* **58**, 1486 (1936).
- [27] H.-A. Loeliger, *Signal Processing Magazine, IEEE* **21**, 28 (2004).
- [28] R. M. Tanner, *Information Theory, IEEE Transactions on* **27**, 533 (1981).
- [29] M. Ramezanali, P. P. Mitra, and A. M. Sengupta, *arXiv preprint arXiv:1509.08995* (2015).
- [30] M. Andersen, J. Dahl, and L. Vandenberghe, *CVXOPT: A Python Package for Convex Optimization* (2010).
- [31] M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory* (Westview, 1995).